**The Dissertation Committee for Sara Stewart certifies that this is the approved version of the following dissertation:**

# Aptamers as cross-reactive receptors: Using binding patterns to discriminate biomolecules

**Committee:**

_____
Eric Anslyn, Supervisor

_____
Andrew Ellington, Co-Supervisor

_____
Scott Hunicke-Smith

_____
Clause Wilke

_____
William Press

_____
Jason Shear

# Aptamers as cross-reactive receptors: Using binding patterns to discriminate biomolecules

**by**

**Sara Stewart, B.S.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**May, 2013**

# Dedication

To my husband, who has supported me in all my endeavors and to my children, the lights of my life.

# Acknowledgments

I'd like to thank my advisors Eric Anslyn and Andy Ellington. It has been an incredible experience working with these wonderful people. No two people could be more different yet work so well together. I am grateful to Nick Christodoulides, Michelle Byrom, Xi Chen, Manny Hughes and Peter Allen for putting up with my questions and helping me get through to nuances of research. I owe my start to Angel Syrett who offered encouragement and pointed me in the direction I am in today. I'd like to thank Scott Hunicke-Smith and all for the people in the next-gen sequencing facility for helping me understand to process and helping troubleshoot my code. To all the wonderful people in the Ellington, Anslyn and Mcdevitt labs, I'd like to say thank you for all the help and support and practice opportunities. Finally I feel I must thank my cat Flame, who helped to write this manuscript by sitting on the keyboard – repeatedly.

# Aptamers as cross-reactive receptors: Using binding patterns to discriminate biomolecules

Publication no._____

Sara Stewart Ph.D.

The University of Texas at Austin, 2013

Supervisors: Eric Anslyn, Andrew Ellington

Exploration into the use of aptamers as cross-reactive receptors was the focus of this work. Cross-reactivity is of interest for developing assays to identify complex targets and solutions. By exploiting the simple chemistries of aptamers, we hope to introduce a new class of receptors to the science of molecular discrimination. This manuscript first addresses the use designed aptamers for the identification of variants of HIV-1 reverse transcriptase. In this research aptamers were immobilized on a platform and were used to discriminate four variants of HIV-1 reverse transcriptase. It was found that not only could the array discriminate HIV-1 reverse transcriptase variants for which aptamers were designed, it would also discriminate variants for which no aptamers exist.

A panel of aptamers was used to discriminate four separate cell lines, which were chosen as examples of complex targets. This aptamer panel was used to further explore the use of aptamers as cross-reactive sensors. Forty-six aptamers were selected from the

literature that were designed to be specific to cells or molecules expected to be in the surface of cells. This panel showed differential binding patterns to each of the cell types, displaying cross-reactive behavior.

During the course of this research, we also developed a novel ratiometric method of using aptamer count derived from next-generation sequencing as a method for discrimination. This is in lieu of the more commonly used fluorescent signals.

Finally the use of multiple signals for pattern recognition routines was further explored by running various models using artificial data. Various situations were applied to replicate different possible situation which might arise when working with macromolecular interactions. The purpose of this was to advance the communities understanding and ability to interpret results from the pattern recognition methods of PCA and LDA.

# Table of Contents

## List of Tables

## List of Figures

# Chapter 1: The Science of Receptors and Their Uses in Chemistry and Biology

## 1. Introduction

The ability to detect something tiny, beyond our scope of vision, is indispensable to the study of the world around us. Early scientists, like the alchemist Robert Boyle, sought, among other pursuits, to identify the composition of common compounds. Boyle is even credited with coining the term "analysis" as a method of detecting ingredients in complex mixtures (Debus, 1962).

In the early days of science, the microscope was the powerhouse of detecting and identifying small targets. Marcello Malpighi is attributed the discovery of many small anatomical structures through his use of microscopy. Henry Sorby developed the science of petrography, the study of rock and mineral characteristics, through his desire understand physical geology (Nutall, 1981). Even in those nascent years, scientists knew there were particles smaller than what could be seen with a microscope. The challenge faced by the early scientist was how to detect those tiny particles.

This challenge was addressed in part by using receptors, molecules capable of transducing a binding event into another form information. Perhaps the most widely recognized class of receptors is biomolecules. For example, the epidermal growth factor receptor, EGFR, is a trans-membrane receptor which binds to a singling molecule from an extracellular source. Upon binding, the extra cellular domain undergoes dimerization resulting in autophosphorylation of two receptor molecules. These phosphorylated

receptors are capable of providing a new binding site for an intracellular molecule, thus initiating a signaling cascade directing cellular behavior. What has fundamentally happened is detection of a binding event and transduced into cellular activity; cell differentiation for example (Fischer et al., 2003).

Receptors however, are not limited to just biomolecules. Any binding event which can be transduced to a different signal can be considered a receptor. Erythrolitmin, a compound found in Litmus paper, is a very early example of a colorometric receptor. In this case, surplus $H^+$ ions protonate the molecule altering the excitation wavelength of the molecule thus resulting in a color shift (Horobin and Bennion, 1972).

Today colorimetric receptors are ubiquitous. The home pregnancy test detects human chorionic gonadotropin (hCG) in urine. In the lateral flow version of the home pregnancy test, the experimental sample flows along a capillary membrane to the reaction zone where a monoclonal antibody to hCG is deposited but not immobilized. If hCG is present in the sample the antibody will bind to the hCG. The sample and antibody mixture then travels down the strip until it encounters an immobilized hCG antibody in the test zone, it will then binds any hCG / antibody units and holds them in an discreet location. Free hCG antibody from the reaction zone continues still further on until it encounters an immobilized antibody that targets the constant region of the anti-hCG antibody in the control zone. The free antibody will then be immobilized in the control zone. The initial hCG antibody from the reaction zone also has a visible latex particle or colloid metal such as gold copper or silver, covalently attached. When the anti-hCG antibody accumulates at the test region due to the presence of hCG or in the control

region because of capture by anti-constant region antibody, a band becomes visible at one or both of those location (Ehrenkranz, 2002).

## 2. Specific vs. cross reactive receptors.

In the model of the home pregnancy test, researchers are striving to develop new and effective receptor systems to detects and or quantify various molecules of interest. There are two general approaches to this: the use of highly specific receptors and the use of cross-reactive receptors. In one approach this specific shape of the target must match that if the receptor. This is called the “lock and key” hypothesis and is illustrated under A in Figure 1.1. Antibodies are frequently considered to follow this model. The alternative approach is the use of cross-reactive receptors. Section B in Figure 1.1 illustrates how this model recognized receptors. Rather than having a single receptor for a target, the target my bind to one or more receptors and each receptor may bind one or more targets.

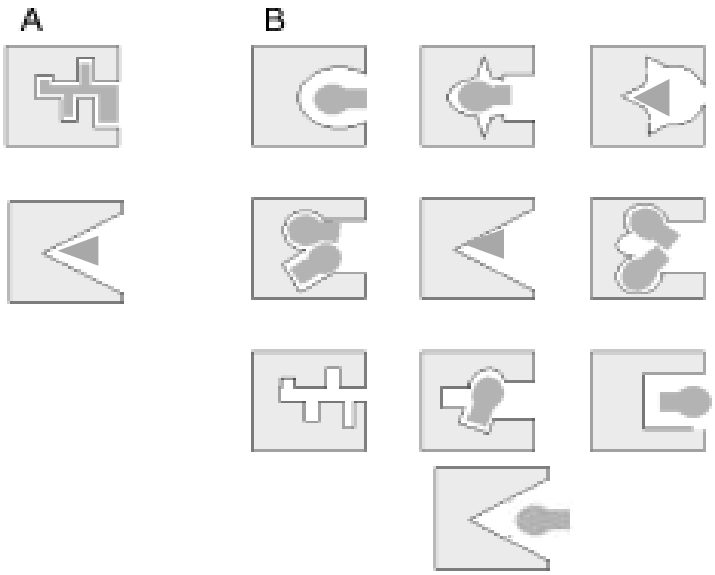



Figure 1.1. Section A represents a "lock and key" model where a single receptor binds a single target. Section B represents a cross-reactive model where receptors can bind more than one target and targets can bind to more than one receptor. Figure adapted from Lavigne and Anslyn, 2001.

### 2.1 Specific receptors

A highly specific receptor is a molecule which will consistently bind to a single target and show very little affinity for non-target molecules. As a general rule of thumb, these highly specific receptors show high sensitivity as well as high specificity though this is not always the case. The GP IIb/IIIa integrin agonist Eptifibatide, is a drug that actively competes with ligands at the site of bridging for platelet aggregation. This particular drug has been shown to bind exclusively to GP IIb/IIIa and not to other integrins. Despite this high specificity, this drug shows a dissociation constant of around 120nM. In contrast, this drug's counterpart Abciximab, has a dissociation constant of around 5nM (Karsten and Weber 2003).

## 2.2 Natural receptors

As previously mentioned the natural world holds many examples of specific receptors. Frequently these receptors are considered “lock and key” type receptors. In this paradigm, the unique tertiary structure of the molecule allows only for interaction with the unique tertiary structure of another molecule (Koshland 1995). Interactions with molecules that are not well suited for the binding pockets of these sorts of molecules are particularly unfavorable; like fitting square peg in a round hole. Antibodies are often considered lock and key type receptor and can show exquisite sensitivity with very little cross reactivity (Branden et al. 1995). It is important to note here that another model of antibody antigen binding exists; the induced fit model. In this model, target binding deforms the receptor structure forcing a “fit” to occur (Koshland, 1995). This is primarily due to residue specific interactions between the antibody and its target. Monoclonal antibodies are routinely generated with dissociation constants in the subnanomolar range, while maintaining their specificity (Vaughan et al., 1996; Hoogenboon 2005; Edwards et al.,2003).

Antibodies are not the only proteins that display lock and key binding behaviors. Lectins are proteins often found in the surface of cells and are hugely specific to sugar moieties (Weis and Drickamer 1996). While there are many characterized lectins, perhaps the most recognizable is the toxin ricin. Ricin is composed of two chains, A and B, with differing function. The A chain is an N-glycoside hydrolase that is well-known to cleave glycosidic bonds in ribosomes (Lord et al 2003). This is the well-known mechanism of toxicity ricin. Ricin chain B, however, shows a characteristic model of

lock and key binding structure. Studies of the bonding affinity of ricin B chain shown the chain has 3 times the affinity for lactose than that of galactose. The position of hydroxyl units on the sugar interact with residues in the protein, specifically Asn46, Lys40, Gln31, and Asn355. Sugars which do not have a conformation amiable to binding to these residues, do not show high affinity for ricin B chain (Rivera-Sagredo et al., 1991).

### 2.3 Unnatural receptors

Receptors with high specificity are not limited to molecules of biological origin. Like ricin B chain, receptors containing boronic acid units are known to specifically bind saccharides and have been researched for over 50 years.

Boronic acid units are arranged with particular orientations in order to match the specific confirmation of the saccharide in question. In aqueous media, the hydroxyl groups of the boronic acids are exposed and accessible to coordinate with the axial or equatorial hydroxyl groups in a saccharide. This method has shown the ability to clearly distinguish between isomers of the same sugar. This is a pinnacle example of a lock and key sensor – only those sugars, whose confirmation is an exact fit for the boronic acid sensor, will bind and therefore be detectable (James et al., 1996).

### 2.4 Non-specific receptors.

Molecule specific detection certainly has great significance in many branches of science; however, there is also a need for methods to detect many different molecules simultaneously. The detection of ions in solution has particular relevance to chemistry

and biology, but due to their small size, it is quite difficult to build a receptor that exclusively detects a single type of ion. Rather, several approaches rely on some differential response in an assay to detect one or several types of ions.

Cation heavy metals can have significant impacts on public health and the environment (Gadd and Griffiths. 1997; Alloway and Ayers 1997). On the other hand, many anions are responsible for acid rain and the decay of manmade structures (Alloway and Ayers, 1997). Rapid assays for the detection of these compounds could play a pivotal role in managing their presence in the environment.

Functionalized gold nanoparticles can be designed in such a way to yield a colorometric response to aggregation. Kim et al., functionalized gold nanoparticles with 11-mercaptoundecanoic acid (MUA). In the absence of a heavy metal the nanoparticles could not aggregate and yield a red color. When a heavy metal cation was added, it coordinated with the carboxylic acid groups between two nanoparticles allowing for aggregation. Once the nanoparticles aggregated, the nanoparticle solution shifted from red to blue.

The Prodi et al. has designed a tripodal ligand, tris[5-(dimethylamino)-*N*-(2-aminoethyl)-1-naphthalenesulphonamide], incorporating the dansyl chromophore, a commonly used fluorescent label (Geddes, 2010). This molecule is capable coordinating a metallic ion resulting in differential fluorescent spectra for several different ions (Prodi et al., 1999).

Another approach, dedicated to the detection of anions, was put forth by the Suslick group. Rather than relying on a fluorescence output, the Suslick group used

anthraquinone, a ubiquitous dye and pigment precursor, as the signal to indicate ion binding. The athraquinone was conjugated to a calix[4]pyrrole, which includes a coordination pocket for accepting anions (Gale et al., 2001). Upon the addition of an anion, a substantial colorometric shift was observed whose magnitude was dependent on the specific ion (Miyaji et al., 2000).

### 2.5 Taste and Smell

Molecules more complicated than simple ions are much more difficult to discriminate with single receptors than ions are. Instead, researchers have turned to the biological senses of taste and smell for inspiration. Both of these senses are examples of chemoreception; the process of transducing a chemical binding event into an interpretable signal. In the case of olfaction, a sensor neuron extends cilia in to the mucus of the nasal passages. Theses cilia express an olfactory odorant receptor protein capable of binding a subset of different chemical odorants such as amino acids, aliphatic alcohols, terpenes, camphor derivatives, and thiols, just to name a few. In the cilia, each cell is capable of detecting just one general group of odorants yet many different members of a group may bind to a single type of receptor (Doty 2001; Holley et al., 1974; Rhien 1983; Gilles and Holley 1984).

The sense of taste, or gustation, is generally thought of as being segregated among 4 (or 5) different tastes: sweet, sour, salty, bitter, and sometimes savory (umami) is also

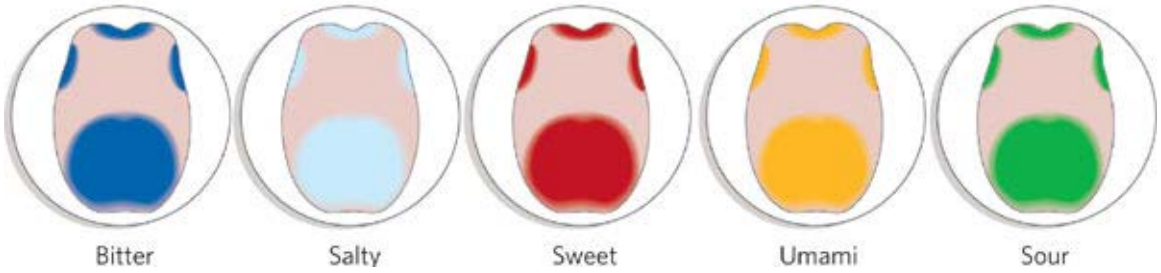



Figure 1.2 Taste locals on the tongue. Figure adapted from Chandrashekar, et al. 2006.

included. This is, however, a somewhat over-simplified view. Each taste bud is comprised of up to 200 separate cells covered with receptors capable of transducing a chemical into the sensation called taste. On the tongue and the inside of the mouth, groups of taste buds are localized into regions where most tastes are detected to greater or lesser degrees. (Figure 1.2) Unlike the taste maps learned in grade school, there is little discreet separation in the where individual tastes are perceived (Chandrashekar, et al. 2006). Within each of these regions there are cells which express certain types of receptors, and which broadly recognize certain categories of tastes. The T1R group of receptors, for example, broadly recognizes "sweet" compounds, not all of which are sugars (Nelson et al., 2001, Zhao et al., 2003). However, one member of this family exhibits strong selectivity for a single compound; the T1R1 receptor seems to bind almost exclusively to monosodium glutamate (MSG) (Li et al., 2001).

Chemoreception in humans and other animals is an example of how a panel of diverse receptors can work in concert to detect and discriminate single molecules, like MSG, or can categorized a group of molecules into a distinct taste/smell.

## 3. Sensor arrays

In order to replicate these powerful discriminator mechanisms of olfaction and gustation, researchers have begun developing panels of receptors. Some receptor arrays have used specific molecules such as antibodies, aptamers or lectins.

### 3.1 Specific arrays

Mor et al. demonstrated that a panel of 4 biomarkers for ovainan cancer, leptin, prolactin, osteopontin (OPN), and insulin-like growth factor-II (IGF-II), are potentially capable of discriminating between clinical and cancer free patients. They used an enzyme-linked immunosorbent assay to detect the various proteins in the serum of sick and diseased patients. With the assay, they were able to correctly classify 95% of sick population with a very low false positive rate. While these results were promising, the assay showed poor prognostic value between different cancer types ( Mor et al., 2005).

McCauley et al. performed a similar analysis using aptamers specifically selected for clinically relevant biomarkers, inosine-5'-monophosphate dehydrogenase (IMPDH), thrombin, basic fibroblast growth factor (bFGF), and vascular epidermal growth factor (VEGF). They immobilized the aptamers on a glass substrate and were able to detect and quantitate levels of thrombin in human serum. They were further able to detect all 4 proteins in bacterial cell lysate, though they were unable to effectively identify the protein concentration, possibly due to degradation of the RNA aptamers (McCauley et al. 2003).

Cell surface saccharides play a role in cell to cell communication, cell adhesion and cell motility. Aberrant expression of cell surface saccharides has been found to be associated with various tumors and may be significant for tumor growth and metastasis. (Orntoft et al., 1990). To assess this idea and to assist in developing lectin based cancer assays, Zhang et al. developed an assay of six lectins immobilized on gold films to study cell surface carbohydrate expression. They found that two cell types, BHK-1 and Caco-2, bound to the array differentially, indicating the presence or absence of certain carbohydrates on the cell surface (Zheng et al. 2005).

While these methods have shown some promise in classifying disease states or identifying cell types, their mechanism of discrimination is not particularly close to that of olfaction or gustation. The previously described methods rely on having (or generating); very specific molecules capable of binding to a very specific target. As mentioned in the discussion of olfaction and gustation, the chemoreception in the human body bind to a variety of different targets with differing levels of affinity.

### 3.2 Cross-reactive arrays

An array that shows behavior more similar to that of natural chemoreceptors is one explored by John Lavigne. In this work, an array was developed by positioning resin beads in micromachined wells formed in Si/SiN wafers. Each bead was functionalized with fluorescein, o-cresolphthalein, alizarin complexome or a boronic acid ester of resorufin-derivatized galactose. These receptors were able to detect variations in pH (fluorescein, o-cresolphthalein and alizarin complexome), $Ce^{3+}$ and $Ca^{2+}$ ions (o-

cresolphthalein and alizarin complexome) or simple sugars (boronic acid ester if resorurin-derivatized galactose). Upon exposure of the chip to solution containing various ions, sugar, and pH concentrations, differential patterns of color response were observed in each of the wells (Lavigne et al., 1997).

Another early approach that mimics the sense of smell is the use of tin oxide receptor for the detection of gasses. This is the principle behind the Warwick nose, one or the earlier attempts at an electronic nose (Albert et al., 2000) In this platform a film of $SnO_2$ inside a ceramic tube, is doped with various precious metals. Upon exposure to various gasses the conductance of the receptors are modulated depending on the metal doped in and the gas itself. This results in a pattern of different conduction peaks over time which are diagnostic of the exact gas compositions (Shurmer and Gardener 1992).

An approach that uses an indicator displacement assay (IDA) was developed to distinguish between ATP and GTP, two very similar molecules. In this study, resin beads immobilized in a Si/SiN wafer were functionalized with guanidinium groups displaying a poly peptide library. A fluorescein was located between the two "arms" of the polypeptides to act as an indicator. When the array was exposed to either ATP or GTP, the fluorescein was displaced from the polypeptide pockets, resulting in a modulation of the fluorescent signal generated by the fluorescein. Using pattern recognition algorithms ATP and GTP were effectively discriminated ( McCleskey et al., 2003).

Dyes that display a color response under various conditions have been used for the detection of organic compounds in water. This is advantageous because many detection strategies for such compound are performed in non-aqueous media, thus they

are not readily generalizable. In this scheme, metaloporphyrins, traditional pH dyes with low aqueous solubility, and solvatochromic dyes were immobilized on a substrate and exposed to solution with various organic molecules. In each cases, a distinctive pattern was generated which was distinguishable with the use of hierarchical clustering (Figure 1.3) (Zhang and Suslick 2005).

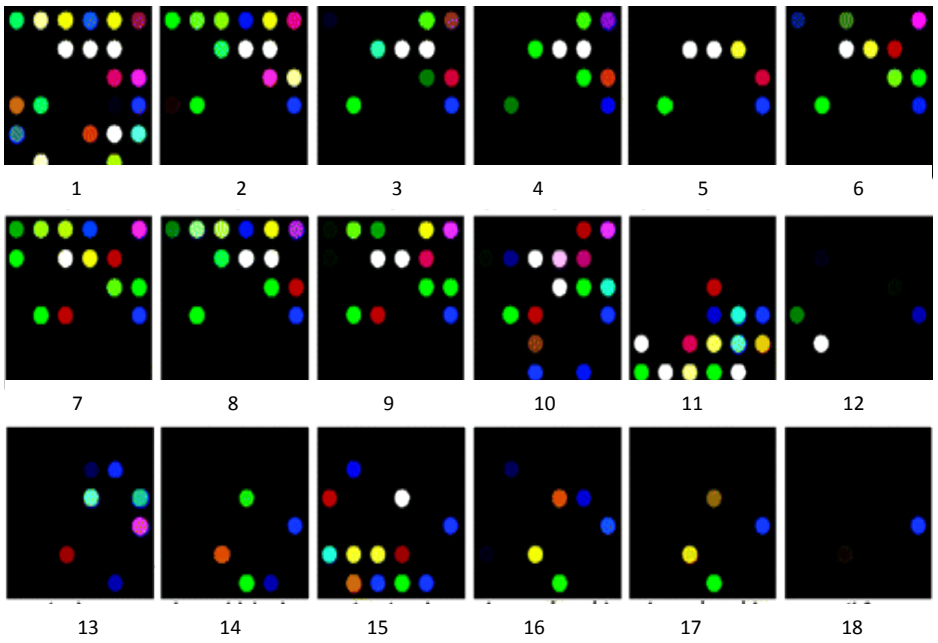



Figure 1.3 Example of differential colorometric patterns generated for 18 different organic molecules. Figure adapted from Zhang and Suslick, 2005.

Small molecules are not the only targets of cross reactive sensor arrays. Rotello has developed a method using gold nanoparticles to distinguish between various cell types including isogenic cells. Each of the nanoparticles was functionalized with a catioinic nitrogen bound to a cyclohexane, a benzyl group or a propanol. These particles were electrostatically coated with a fluorescent polymer and were used to discriminate

cells. Each cell showed a different affinity of each type of nanoparticle resulting in a different fluorescence profile. Using a canonical discriminate analysis, the Rotello group was successful in grouping similar cell types, though in many cases there was very little difference between the lines.( Bajaj et al., 2009).

### 3.3 Cross-reactive arrays based on biologic and quasi-biologic molecules

Of all of the platforms previously describe nearly every one relied on non-natural synthetic molecules. Antibodies, lectins, and aptamers have been used in array platforms, but they have very little utility in cross-reactive systems by virtue of their high levels of selectivity for target molecules. However, there is growing interest in biological molecules for use in differential sensing routines. It is not uncommon for non-biological receptor molecules to be complicated to synthesize. While there has been great strides in rational synthetic receptor design, it is still difficult to predict (and screen) the behavior of synthetic molecules. Instead, biologic molecules may offer improved insights into receptor target interactions. There has been a wealth of research into the structure and function of biologic molecules and with the advent of high throughput methods such as next-generation sequencing, the amount of available knowledge is expanding rapidly (Howe et al., 2008). By exploring novel application for biological molecules as differential receptors, there is the potential to simplify the creation of receptor arrays while simultaneously increasing their repertoire of targets.

The endogenous receptors found in the nose and on the tongue could potentially be interesting targets for use in cross reactive sensing platforms. However, at this time,

the precise nature of the receptors are only just coming to light. This is not to say that there are no other proteins which could be used for differential sensing. Albumin is the most abundant plasma protein and is involved in the transport of many compounds including ions, hormones and various other hydrophobic compounds (Peters, 1995). It has been postulated that the extreme high level of cross reactivity displayed by albumin could be exploited to discriminate molecules in vitro.

Adams and Anslyn exploited the use serum albumin (SA) as a method for discriminating terpenes (hydrophobic molecules present in many varieties of plants) in perfume samples. The fluorophore PRODAN was used as an indicator for the assay. It remained absorbed in the hydrophobic pockets of the SA, along with a hydrophobic additive, until it was displaced by a terpene in ethanol or as part of a perfume sample. The fluoresces of the sample was modulated differentially based on the terpene (Adams and Anslyn 2009) (Figure 1.4).

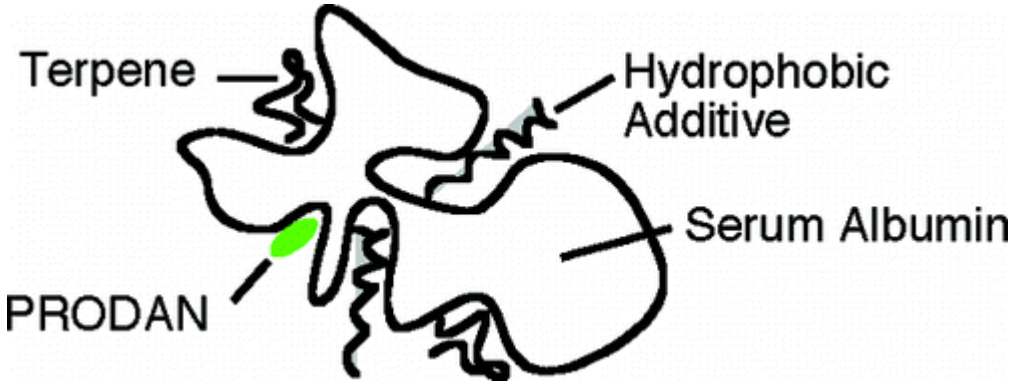



Figure 1.4 Representation of how various molecules non-specifically bind serum albumin. Figure adapted from Adams and Anslyn, 2009.

While antibodies and lectins are complex, relatively large molecules comprised of many amino acids, peptides are small structures that can be synthesized de novo and can be bound by various proteins. The immobilization of peptides on arrays has been researched for some time. However, this method is typically used for identifying protein substrates and epitope mapping (Uttamchandani, et al., 2008). Because of the small size of peptides, they show more cross reactivity than full sized proteins. Robinsen et al. used arrays comprised of peptides, proteins and oligonucleotides to characterize and distinguish several class of autoimmune disease from healthy individuals (Robinson et al., 2002).

Similar to peptides, peptoids are oligo N-subsitited glycines that are synthesized to express side groups on the N of the peptide backbone rather than just the α-carbon. The side group also need not be a natural side group. The Kodadek group explored using as array of 7680 peptoids immobilized on a glass substrate for protein discrimination. They found a unique fingerprint for each of the proteins used: ubiquitin, maltose binding

protein, goat ant-moujse IgG, and anto-glutathione transferase. While limited in scope, their work has shown the utility of peptiods for biomolecular discrimination. (Reddy et al., 2005)

Aptamers were developed in the late 80's early 90's as potential alternatives to antibodies. They were inspired by naturally occurring catalytic ribozymes and the protein RNA complex; the ribosome (Ellington and Stozak 1990; Tuerk and gold 1990). Today aptamers are a popular topic of research and have been generated to target a wide range of substrates such as ATP; aminoglycosides, metallic ions, and porphyrins ( 2001; Sazani et al., 2004; Walter et al, 1999; Li et al. 1996; Zhang et al, 2011). For the most part aptamers are considered to be highly specific molecules, but there is evidence that this may not always be the case. Some aptamers have been designed to express cross-reactivity such as the "toggle" aptmer for thrombin. This aptamer was specifically selected from a standard pool of aptamers to bind both human and porcine thrombin (White et al. 2001).

Other aptamers have been found to bind to a family of proteins. This is the case for RNA-1 an aptamer put forth by Weiss et al. This aptamer was derived from a pool of aptamers designed to target TFIIIA. However, rather than showing specificity to just TFIIIA as its counterpart RNA22 from the same panel did, this aptamer showed high sensitivity to all tested zinc finger proteins (Weiss et al., 2010).

Even aptamers purported to be very specific to their targets have been shown to have at least some off target cross-reactivity. A study performed by Na Li showed that

some aptamers, one to CEM cells (Shangguan et al., 2006) in particular, show binding to unexpected targets (Li et al., 2009).

Some aptamers, however have been designed to show broad cross-reactivity to several different targets. Inspired by previous work on the detection of molecules in a three-way junction binding pocket of an oligonucleotide (Kato et al., 2000), the Stojanovic group developed a thiol modified aptamer capable of discriminating cocaine and various other hydrophobic ligands. Like the previously discussed work by Adams, this method relies on the displacement a fluorophore in the three way junction by the molecule of interest. Upon binding, modulation of fluorescence identifies the exact compound located in the three-way junction. (Stojanovic et al., 2003).

This aptamer work has led to the idea that aptamers may show much more cross reactivity than previously thought. This manuscript will discuss two separate studies in aptamer cross reactivity and the significance to aptamer cross-reactivity to biomolecular sensor arrays. The second chapter will focus on an aptamer microarray technique employing aptamers specifically selected for wild type HIV-1 RT and a mutant variant M3. It will show that this array of aptamers can not only discriminate between the proteins the aptamers were selected for, but also proteins the aptamers had never been exposed to. Chapter 3 will discuss a panel of 46 aptamers selected from the literature known to bind various molecules expected to be found on the surface of cells. These 46 aptmers will be used to discriminate between 4 separate cells lines. A novel ratiometric method of using aptamer count derived from next-generation sequencing will be used as a signal in lieu of a fluorescent label. Chapter 4 will discuss the computational

techniques employed throughout these two studies and modeling work done in order to better facilitate the community's understanding of the pattern recognition techniques used here.

# Chapter 2: Identifying Protein Variants with Cross-reactive Aptamer Arrays

*This chapter is derived from "Identifying Protein Variants with Cross-Reactive Aptamer Arrays." by Stewart, S., Syrett, A., Pothukuchy, A., Bhadra, S., Ellington, A., & Anslyn, E. ChemBioBhem, 12.13(2011) 2021-2024.*

## 1. Introduction

Aptamers are structured DNA or RNA molecules selected from random sequence pools (Ellington and Stozak, 1990; Tuerk and Gold 1990). Through an iterative process, aptamers are generated that can show a range of sensitivities and specificities to target analytes, and that are frequently able to cross-recognize protein variants. For example, different aptamers selected against single amino acid variants of bacteriophage MS2 coat protein (Hiaro et al., 1998) were found to exhibit a range of activities, from binding to only one protein variant, to avoiding only one variant, to generally binding each of the protein variants. Recently, aptamer microarrays have been used by biological researchers to identify the presence of relevant biomarkers from a fluid of interest (Miller et al., 2003). In this regard, aptamers have previously been used for biosensing (Kirby et al., 2004; McCauley et al., 2003; Nieet al., 2007; Stadtherr et al., 2005; Li et al., 2007; Stefaha al., 2009). In general, theses arrays rely on aptamers with a high affinity to a single target and very low affinity to non-targets (Conroy et al., 2009; Luppa et al., 2001; Goodey et al, 2001; Haab et al., 2001). In order to create such aptamers, a substantial amount of time and effort must be invested to identify aptamer candidates. Our hypothesis was that an array of potentially cross-reactive aptamers might be able to more broadly recognize a series of protein variants than a similar set of highly specific

receptors (Lavine et al., 2001; Dickenson et al., 1996; Albert et al., 2000, Lewis et al., 2004). Aptamers that recognize families of proteins rather than a single member have been generated. One such example is RNA-1, isolated from a selection against *Xenopus*. Strikingly, this aptamer was found to bind broadly to a variety of different zinc finger proteins (Weiss et al., 2010). Stojanovic has shown aptamers' affinity for a target can be perturbed by introducing single mismatches near the binding pocket and varying the position in the aptamer to which a fluorophore was conjugated (Stojanovic et al., 2003). They were able to generate a panel of modified aptamers capable of discriminating between cocaine and various cocaine analogs. Pei et al. describes an assay in which systematic variations in aptamer sequence, with both natural and unnatural nucleic acids, were used to discriminate 12 different alkaloids (Pei et al., 2009). Rather than relying on systematically varying a single aptamer, we hoped that the aptamers normally generated by *in vitro* selection would serve as a set of semi-specific receptors.

This principal of using an array of semi-specific receptors has been extensively exploited using synthetic receptors for a wide variety of compounds, thus eliminating the need for a unique receptor for each target (Shangguan et al., 2008; Fitter et al. 2004). Suslik et al. has developed a sensor array using metalloporphyrin dyes coordinated with different ions (Suslik et al., 2004). These dyes readily undergo a colorimetric shift upon binding of a ligand. The group was able to generate unique patterns of binding for 14 different compounds in vapor (Rakow et al., 2000). The Anslyn group popularized the use of indicator displacement assay (IDA) through their use of various cross-reactive sensors based on peptide-based receptors derived from combinatorial library synthesis

(reviewed by Nguyen, 2006; Umali et al., 2010). We have illustrated this approach by using a hexasubstituted aryl core decorated with guanidinium groups which were appended to combinatorially-generated tripeptide arms. This approach has also been to discriminate protein targets. We generated a library derived from appending variable peptide arms to a tridentate core, along with different metallic ions and counterions chelated into the tridentate core. An 18-receptor ensemble was used to effectively discriminate α-neurokinin, substance P, and tachykinin (Wright et al., 2007). This approach has been further expanded to classify complex targets. In this example small peptidic sensors, 6 to 9 amino acids in length, were used to discriminate flavonoids and red wine varietals (Umali et al., 2011).

Previous work performed by Drs. Angel Syrett and Na Li generated a number of aptamers with various affinities to wild-type HIV-1 reverse transcriptase and a mutant variant of HIV-1 reverse transcriptase called M3 (Li et al., 2009; Syrett 2010). We wished to use these aptamers in differential sensing protocols to distinguish between wild-type and several drug resistant variants of HIV-1 reverse transcriptase. The ability to effectively discriminate various mutant strains of HIV-1 reverse transcriptase could have significant implications on the appropriate course of treatment. Certain mutations, such as substitutions at RT codons 41, 67, 70, 210, 215, and 219 allow reverse transcriptase to resist inhibition from nucleic acid analogs such as AZT, a common drug used to treat HIV infection (Coffin et al., 1997). HIV infected infants can have a 50% mortality rate before the age of 2 if left untreated. Swift, appropriate medical intervention can reduce the mortality of these children significantly, but only if begun early in the

infection (Pennazato et al., 2012). While some current assays can detect drug resistant mutations, they are time consuming, taking between 1 and 4 weeks to complete (Sen et al., 2006). An immobilized aptamer array might be capable of directly identifying the drug resistance profiles of circulating viruses.

## 2. Experimental methods

### 2.1 Preparation of aptamers

Aptamers were selected against either wild-type (WT) reverse transcriptase (RT) or a drug resistant variant (M3) (Roland et al., 2009; Burke et al., 1996; Tuerk et al., 1992). All DNA and primers were acquired from Integrated DNA Technologies (Coralville, IA). LNAs were obtained from Molecular Probes (Eugene, OR). The forward primer sequence was

5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGGGG-Constant1-3' where the underlined region corresponds to the region that will be complementary to a Cy-5 labeled LNA probe, and the sequence of the constant region can be found in table S1. The purpose of the Cy-5 labeled LNA was to allow for approximate quantitation of the amount of aptamer bound to the slide. Constant1 was the region complementary to the primer region found on the experimental aptamers. The reverse primer was 5'-TTCTCGTGATGTCCAGTCGC-Constant2-3' where the underlined region corresponds to the sequence of a biotinylayted LNA anchor. The purpose of the biotinylated LNA was to bind the aptamer complex to the Strepavidin slide. Constant2 is a region

complementary to the reverse primer region of the experimental aptamers. The general structure of the aptamer complex was LNAProbe-Constant1-Active region-Constant2-LNA anchor. Figure 2.1 shows a graphical representation of how the biotinylated anchor, aptamer, and labeled probe come together on the slide to form the sensor.

It should be noted that there are different constant regions present in the aptamers, derived from different initial selections, as detailed in was a Table 2.1, located at the end of this section.

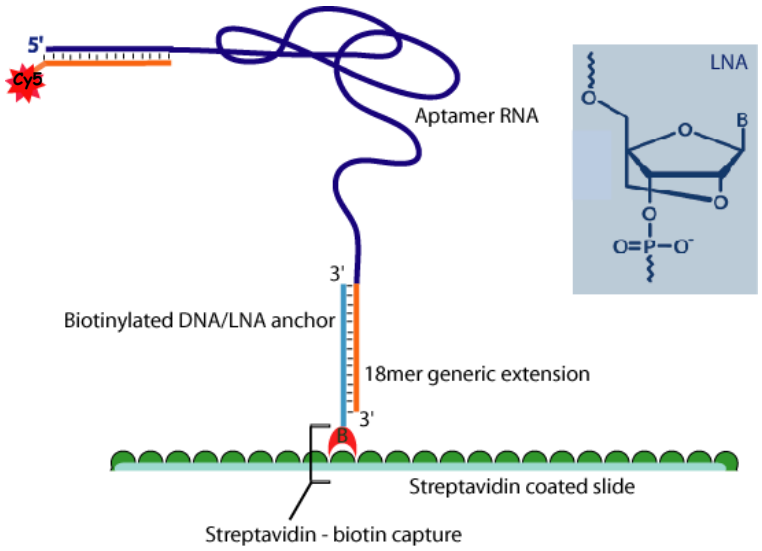



Figure 2.1. A biotinylated LNA anchor complementary to the 5' end of the aptamer, securing the aptamer to the Nutravidin coated slide. A second LNA conjugated to the 3' end of the aptamer acts as a probe for detection of bound aptamer. Figure courtesy of Angel Syrett.

RNA aptamers were generated from DNA templates by Epicenter's (Madison, WI). Ampliscribe kit RNAs were purified on an 8% PAGE gel followed by elution, ethanol precipitation, and suspended in water. RNA concentrations were measured using a NanoDrop at A260 nm. In order to generate the full construct in preparation for printing, each aptamer was hybridized at the 5' end to the Cy-5 conjugated LNA probe and at the 3' end to the biotin-conjugated LNA anchor. The hybridization of 2 μM aptamer to 5 μM of each LNA was carried out in 1X reverse transcriptase buffer (RTB; 100 mM NaCl, 20 mM Tris-HCl, 5 mM $MgCl_2$, 1 mM DDT) at 70 °C for 3 min, followed by a drop to 4 °C at 0.1 °C/sec. After hybridization, 3 μl of 50% glycerol was added to 30 μl of the aptamer solution.

| Name | Sequnce | Consants used 1, 2 | Protein Selected For |
|---|---|---|---|
| Costant R | AAGTGACGTCTGAACTGCTTCGAA | | |
| Constant B | GGGAAAAGCGAATCATACACAAGA | | |
| Constant L | GGGTTACCTAGGTGTAGATGCT | | |
| e13 | AGAGCGUCGAUGAUAGUUUGGAAGGCGUCC | L,R | Wild-Type |
| e18 | CUGGGCUCUAGAAGUUCUGCAACUUCGAAU | L,R | Wild-Type |
| e19 | AUCCUGGAACGUACAGCCGGACGUAUAAAU | L,R | Wild-Type |
| e24 | GUACUGGAGCGUCGACAACGGCUUCGUAGC | L,R | Wild-Type |
| e31 | ACAUGUAGAGCGUCGACGUGCUACCACGUU | L,R | Wild-Type |
| e32 | GUGGCGUAGAACGUUCUGUGGCCUUCGAAC | L,R | Wild-Type |
| e37 | CACAUAGCCGCCAGAAACGUUCCGUCACCG | L,R | Wild-Type |
| m306 | UAACAUAGUUCUCGAAGUCCUUUGUAAGUGGGCUUCCAGAGCUACCAUUU | L,R | M3 |
| m307 | UAACAUAGUUCUCGAAGCCCUUUGUAAGUGGGCUUCCAGAGCUACCAUUU | L,R | M3 |
| m318 | GAUAUUCGCAUUGGCUUCGUCUCAUUUUUAAAUUUUCUGCGGAUCGCAAGC | L,R | M3 |
| m321 | GGGUUACCUAGGUGUAGAUGCUCACUUACACAAACCUGUAGUUUUAUGCUUCACAU<br>AUCGCACAGGACGCUUAAGUGACGUCUGAACUGCUUCGAA | L,R | M3 |
| m324 | UCAGGAAACAGCUAUGACCAUGAUUACGCCAAGCUUGGUACCGAGCUCGGAUCCAC<br>UAGUAACUGCCGCCAGUGUG | L,R | M3 |
| m326 | UUUCAGAUUCAUCUAUAAGAUCNGACAACCGGUUNNUUAGUNGNCNCGAUU | L,R | M3 |
| 12.01 | UCAGAUUCAUUUAUAAGAUCCGACAACCUGUCUAUUAGUUGCGCCGAUU | L,R | M3 |
| c1 | UGCUGACGGCGUGAUAUAAUAAUAACCUACAUGCUCUUCGGCCUUCGGAAGAAUCA<br>CGCUAGAGUAGCGAGUCAGCGAACACGCAUCUAUGACCAGGUUAU | L,R | M3 |
| c2 | CUGCCAUAAUAUUUCAAACUUUCUAAUCAUUUUCCUGUUCUUAUUCUGCCGGAAUC<br>UUUUCACGCAUUUUCUCUUAUUCCGGCUCCUG | L,R | M3 |
| c3 | UGAUGCUGUAAUUGGCAUCCAAUCUGCGCCAACAGCCUUUUUAGUCGUUACUUGAG<br>UAGCGUGCGAUUUGUGAAACCGCUCAAUACGACUAGGUAGAG | L,R | M3 |
| c5 | AAUGUAUAAGGAGAGCUCCGUAAGGACUGUCCCGCCUAUGACUGUCGACUCAGGUC<br>GUAAAACUGCCUCUCAGGAUACAAGUAUCACCUAGUGUUGGCA | L,R | M3 |
| c9 | ACCCCGACUCGUGUUAGCACAUAGGACGGGUGCUGACCCCCUGCAACUAACUCCUC<br>AAGGGAUACAAGCGCGGAGUCUGGCACUAACCAAGUACCGGG | L,R | M3 |
| c20 | AUUUUCGUUGAUGCGGAUACCUGAACCAUGGACUUGGUCGACCUCAAUGUAAAACA<br>GGUGAAUCGCUAACUUAGGAUUGCACCACGCUAUGAGAGGUCGUC | L,R | M3 |
| c24 | UGCCGCAGCCUCUAAGUGUAAGCAUCAAUCUGGGUAAACCCGUAGACGAAAUAUGG<br>CUCUCGAUCCGUUAGGGAGGCAAGACUACGGACCCAGUUUUGCG | L,R | M3 |
| c30 | AACACAGGGUGGCUAACCCCUCAAAGUAGACUGUUAAGGACUUCACGGUCGUAAAA<br>CCAACUCCACGGGUGUCUCCUGAUGGCGUUG | L,R | M3 |
| c31 | GCAAUUUAAGGCAUGAAAAUCGUACACAUAACAUGGCCUAUCUACGACCGUACGUC<br>AAUCACUGUCAACCCACAACUGUGGUUACUUGGUUUCC | L,R | M3 |
| c35 | AUAACAAUAAUAAUAUUCGUACAGUGCUGCCGCCUUUGAAGAUAACCGCUCACUUG<br>UAAGCGGGGAAGUUCGGCUACGGGGUUAGCAAGGUGCGG | L,R | M3 |
| c39 | CUAGCCUACGCUUUCUAUUAUAACAUAAACAUAUACCACGUGUAACUUGGGCCAAA<br>CCCUAGUCUUUGUCCGUUUAGGCCCUCUGGUGCAGCUCCG | L,R | M3 |
| 70.01 | AAGAAAUAUCCGUUACCAUUCGGGAAAAAAUGGGUUGGUG | B, R | Wild-Type |
| 70.04 | AUGCGACUUCCAAAAGAGAUCCAGGGAGCAGGCGCACUGGGAGAAAACU | B, R | Wild-Type |
| 70.08 | AAAAGAGAUCUGGCAGUGUCACAACCAGGAAAAAGACACGACGAAC | B, R | Wild-Type |
| 70.12 | ACAAGAUAUCCGAGCCAAAACGGGAAAAGAUGGGAAAAAU | B, R | Wild-Type |

Table 2.1. List of constant regions and variable regions for each aptamer used.

## 2.2 Preparation of mutant reverse transcriptase

M3, M5 and M9 HIV-1 RTs were prepared as previously described (Hou, 2004; Li 2009) with some minor modifications and obtained from our collaborators at Accacia, L.L.C. (Austin, TX). Briefly, two subunits, p51 and p66, were cloned into the vectors pET30 and pET21a, respectively. The subunits were expressed in BL21(DE3) codon plus RIL *Escherichia coli* strain (Novagen, Gibbstown, NJ). Cells expressing each subunit separately were disrupted using a flow-through, high-pressure homogenizer and centrifuged at 28,000*g* for 45 min. at 4 °C. The supernatant was loaded onto a DEAE-Sepharose column (Promega, Madison, WI) equilibrated with buffer A (50 mM Tris-HCl pH 7.9, 60 mM NaCl, 8% glycerol v/v, 1 mM DTT). The flow-through was supplemented with NaCl to 500 mM and imidazole to 10 mM, and loaded onto a Ni-NTC column (Qigen, Valencia, CA) that was equilibrated with buffer B (50 mM Tris-HCl pH 7.9, 500 mM NaCl, 8% glycerol v/v, 1 mM DTT) and imidazole to 10 mM. The column was washed with 12 column volumes of buffer B supplemented with NaCl to 1M. This was followed by 12 column volumes of buffer B supplemented with imidazole to 5 mM and 12 column volumes of buffer B supplemented with imidazole to 10 mM. Each of the RT variants was eluted with buffer B supplemented with imidazole at 40 mM, 60 mM, 100 mM, and 200 mM (used sequentially). Each fraction was assayed for purity on an 8% SDS-polyacrylamide gel and dialyzed overnight into 2X storage buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 20% glycerol v/v, 1 mM DTT) using 3-ml Slide-a-lyzer cassettes with a 3500 molecular weight cut-off (MWCO) (Pierce, Rockford, IL). All of the proteins used, whether purchased or prepared by collaborators, were tested for activity via a

| AA number | WT | M3 | M5 | M9 |
|---|---|---|---|---|
| 41 | M | L | M | M |
| 44 | E | D | E | E |
| 67 | D | N | N | D |
| 69 | T | D | D | T |
| 70 | K | K | R | K |
| 75 | V | V | V | I |
| 77 | F | F | F | L |
| 116 | F | F | F | Y |
| 118 | V | I | V | V |
| 151 | Q | Q | Q | M |
| 184 | M | V | V | M |
| 210 | L | W | L | L |
| 215 | T | Y | Y | T |
| 219 | K | K | Q | K |

Table 2.2. Location and Amino acid identity for proteins used. (Rhee, 2003)

polymerase extension activity assay. Briefly, 5 nM of a 100bp RNA template:primer mix was incubatedwith 50 µM of dNTPs and RT (15, 50, 500 nM) at 37 $^{o}$C for 10 min. Reactions were quenched with EDTA. The products were run on an 8% denaturing polyacrylamide gel and activity was assessed by determining that the products

were of the correct length. Protein concentration was determined using the Bradford assay.

### 2.3 Preparation of slides

Neutravidin slides were obtained from Pierce Biotechnology (Rockford, Il). The aptamer preparation was printed on Neutravidin-coated slides by a capillary action arrayer at 75% humidity. 10 µl of each aptamer was loaded into a distinct set of wells in a 384 well plate. The DeRisi capillary action arrayer is equipped with a tandem print head to allow for the printing of up to 16 arrays on a single slide. Each aptamer was drawn into the pinheads through capillary action and spotted either in triplicate or in sextuplets. Between the printing of each aptamer the pins used for printing were washed in a sonicator filled with 1X SSC buffer (150 mM sodium chloride, 15 mM sodium citrate). The pins were dried for 10 seconds 3 times under vacuum prior to retrieval of the next aptamers. After printing, each slide was incubated in the humidity chamber for thirty minutes to allow for maximum capture of the biotinylated aptamer complex by the streptavidin on the slide. After incubation, 150 µl of Blocking Buffer (1:1 RTB, Roche western blocking buffer) was added to each well and incubated overnight at 4 ºC. After incubation, the buffer was removed and the slides were washed 3X with RTB. As a result of the printing method each slide was printed with 16 identical arrays. In order to isolate each array from the others, FlexWell partioning pads (Grace Biolabs) were applied to the slides. The precise location of each aptamer and the treatment of each well

on each slide varied from experiment to experiment; this information will be found in later sections.

### 2.4 Slide assay

In order to detect the HIV-1 reverse transcriptase variants, a sandwich assay strategy was used. Wild-type reverse transcriptase was acquired from Ambion (Austin, TX). The mutant variants were produced in-house though a collaboration with Accadia (details are found in the section titled “2.2 preparation of mutant reverse transcriptase”). Rabbit anti-RT was use as the primary antibody and obtained from NIH AIDS Research and Reagents program (Stuart Le Grice). Cy-5-labeled anti-rabbit was used as the secondary antibody to detect the rabbit anti-RT (Amersham). A blocking buffer consisting of a propriety protein blend including milk casein was obtained from Roche (Indianapolis, Il).

In order to limit nonspecific binding, each slide was blocked overnight with 1X Blocker (1:9 Roche buffer:RTB). An aliquot of 850 pM reverse transcriptase in RTB was prepared for each of the mutants. 100 µl of the protein solution was incubated in each well for 30 min. In order to limit evaporation of the protein solution during this step and each subsequent step, a plastic film was applied to the FlexWell partioning pads and the slides were kept in a humidity chamber of at least 70% RH. The protein solution was removed from the wells and the slides were rinsed 6x times with 1X Blocker. 100 µl of a 1:10000 dilution in 1X blocker of rabbit anti-RT antibody was incubated in each well for

2 hr. The antibody solution was removed from the wells and then washed 6x times with 1X Blocker. Finally 100 µl of a 1:10 dilution of goat anti-rabbit antibody was incubated in each well for 1 hr. After removing the secondary antibody solution the slides were washed for 5 min. in RTBT (1:1 RTB: 10% Tween-20), then 5 min. in RTB, and finally 5 min. in Nanopure water and allowed to air dry.

### 2.5 Data analysis

Immediately after being treated with the proteins, the fluorescence intensity was measured with a GenePix 4000a fluorescence imager. Foreground and background (Figure 2.2) values for Cy-5 (bound to the secondary antibody) and Cy-3 (bound to each aptamer) were measured. The rationale for measuring the relative abundance of both the

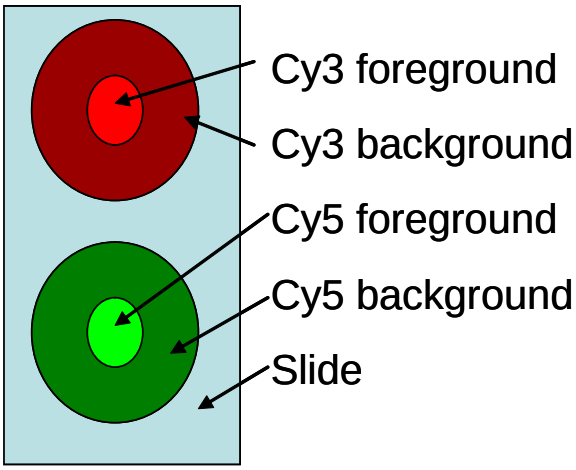



Figure 2.2. For data analysis the foreground signal represented the spot where the aptamer material was printed. Background was a circular region surrounding each spot.

protein and aptamer concentration was to normalize the amount of protein captured to the amount of aptamer bound to the slide. For each spot and each channel, the background was subtracted from the foreground. Spots where the Genepix automated spot locator was unable to find signals were excluded from analysis. Spots where the background intensity exceeded the foreground intensity in a single channel were excluded as this would result in a negative signal value. Each slide was median-normalized using local background intensity and positive and negative control spots. The average of the replicates in the Cy3 and Cy5 channel were calculated; if, for a single group of replicates, there were no acceptable spots, that value was flagged as missing. Depending on the number of missing values for that aptamer or well, the data was either estimated or excluded. The number of missing data points was calculated for each well on a slide, and if more than 15% of the data was missing, the well as a whole was excluded from analysis. The $\log_2$ ratio of Cy5/Cy3 was calculated and was passed to XLStat for PCA (Pearson) or LDA analysis. Missing values were estimated by the XLstat package through the K nearest neighbors (KNN) method based on values observed in other wells from the same treatment group (Troyanskaya et al., 2001). Background subtraction and normalization was carried out by the *marray* microarray statistics package (Bioconductor) (Gentlemen et al., 2004) of R. PCA and LDA were carried out by XLStat, and 3D plots were generated in the commonly used statistics package SPSS.

## 3. Rationale for use of LDA

Linear discriminant analysis (LDA) is referred to as a supervised learning method, as opposed to principal component analysis, which is an unsupervised method. It is referred to as supervised because the method takes in to account group membership when building the model. Fundamentally the LDA algorithm rotates the data in n-dimensional space such that the within-group scatter (if scatter is the parameter being used) is minimized and the between-group scatter is maximized (Equation 2.1).

$$criterion = inv(S_w) \times S_b$$

Equation 2.1 The criterion for LDA analysis.

This criterion is ultimately maximized as an eigenvalue problem (Martinez et al., 2001). Early approaches to solving eigenvalue problems involve diagonalization of a matrix to find the underlying eigenstructure. This is the case for Eigenvalue Decomposition (EVD) equation 2.2, where A is a symmetric matrix (typically a covariance matrix), E is the matrix of eigenvectors and D is the diagonalized matrix of eigenvalues (Shlens, 2009). This approach, while valid, is very computationally taxing and thus is no longer preferred.

$$A = EDE^T$$

Equation 2.2. Formula for eigenvalue decomposition.

However, other methods for solving for the underlying structure do exist. Singular Value Decomposition (SVD) is one of the more commonly used approaches and is quite similar to EVD in some ways. The general form of SVD is found in equation 2.3.

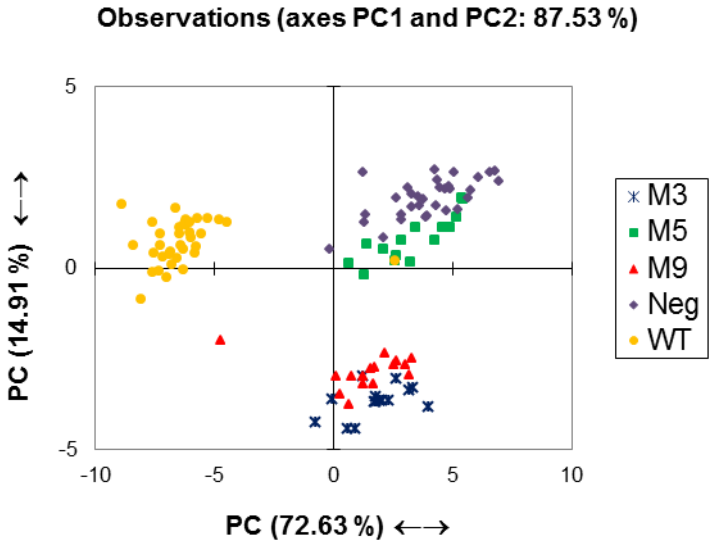
$$X = U\Sigma V^T$$

Equation 2.3 Formula for singular value decomposition.

Like EVD, V is a matrix of eigenvectors and Σ is a diagonalized matrix. In this case however, U and V are termed the left and right singular vectors and the diagonal of Σ are the singular values which are the square roots of the eigenvalues. Unlike EVD which requires a symmetrical matrix to compute, SVD has no such limitation. Thus the X in SVD is the original data itself and creating a covariance matrix is unnecessary.

When using LDA one must be wary of the risk of artificial segregation induced by the use of supervised learning methods. Since information regarding class membership is used to build the model there is a potential for the model to simply predict itself. A model which is not at risk for spurious segregation is Principle Component Analysis (PCA). This method, like LDA, is an eigenvector problem, but it does not use class membership to build the model. Instead, this method finds the axis with the greatest variance. This, of course, makes the assumption that variance in the variables is important information and is relevant to the differences in the groups.

It was decided to first determine if any differential signal could be detected through PCA. In the plot (Figure 2.3) we observed a separation of the wild-type protein from the mutant proteins and negative controls across the first principal component,



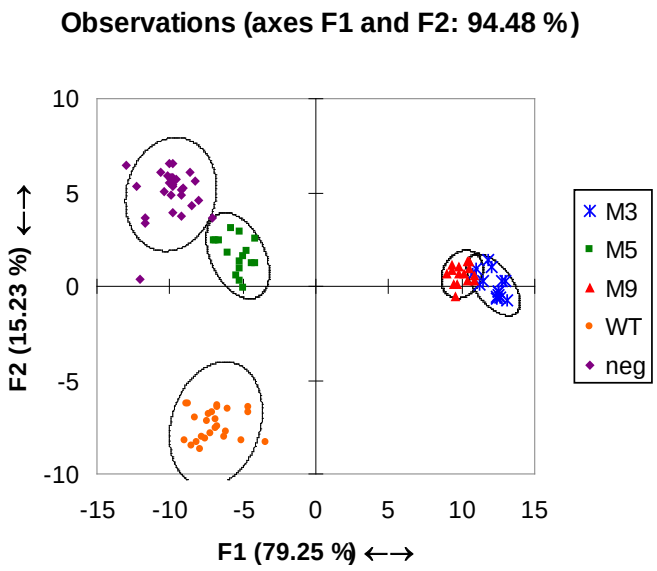

| | M3 | M5 | M9 | WT | Neg |
|---|---|---|---|---|---|
| M3 | 0 | 1.869 | 0.343 | 2.578 | 2.344 |
| M5 | 1.869 | 0 | 1.581 | 2.339 | 0.478 |
| M9 | 0.343 | 1.581 | 0 | 2.291 | 2.059 |
| WT | 2.578 | 2.339 | 2.291 | 0 | 2.561 |
| Neg | 2.344 | 0.478 | 2.059 | 2.561 | 0 |

Figure 2.3. PCA of unnormalized data from 30 aptamer set, 87.53% explained and the Mahalanobis distance.

while the mutants and negative control formed two separate groups across the second component. From the PCA plot, it appears that there is indeed separation of the various reverse transcriptases, albeit very poor separation in the cases of M5/neg and M3/M9. Because was there appeared to be at least some grouping based on treatment group and there does not appear to be any distinct artifact discrimination, it was decided that LDA could be used to improve the segregation of the data.

### 3.1 Rationale for normalization

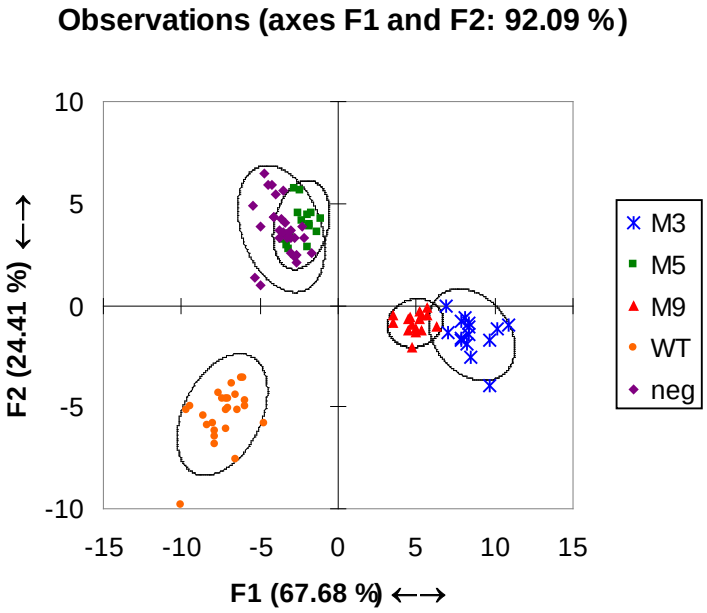
An essential part of developing any model is to correctly account for undesierable variance. For these experiments, one major source variation is the fact that different slides were used. Furthermore, it is possible that the position of each well on the slide could alter its characteristics. In order to account for this, a number of wells on each slide were treated as positive or negative controls. These controls should show the behavior across all slides. It was also assumed that the median value in the Cy5 and Cy3 channels should be the same for a single treatment group. To eliminate extraneous variation, each slide was median-centered, such that the median values of signal intensity were the same for each well treated with the same antibody on a single slide (within-slide). This was followed by a scale normalization of the entire set of slides using the positive and negative control wells (between-slide). This made the median absolute deviation (the median of the absolute deviations from the data's median) for all the control wells equal. Figure 2.4 shows the unnormalized LDA plot for the full set of 30 aptamers. Figure 2.5 shows LDA of the set after only “within-slide” normalization. The unnormalized data




| | M3 | M5 | M9 | WT | neg |
|---|---|---|---|---|---|
| M3 | 0 | 333.604 | 20.175 | 419.382 | 516.726 |
| M5 | 333.604 | 0 | 256.243 | 104.177 | 71.381 |
| M9 | 20.175 | 256.243 | 0 | 352.367 | 426.451 |
| WT | 419.382 | 104.177 | 352.367 | 0 | 162.482 |
| neg | 516.726 | 71.381 | 426.451 | 162.482 | 0 |

Figure 2.4. LDA of unnormalized data from 30 aptamer set, 94.48% variance captured, and Mahalanobis distance.

appears to capture 94.48% of the variance. However there are several points that cluster with the incorrect group, leading to a confusion matrix (a matrix expressing the probability of correctly classifying all samples) result of 99.4%. When only within-slide normalization is performed, the confusion matrix result is reduced to 98.75%. When both




| | M3 | M5 | M9 | WT | Neg |
|---|---|---|---|---|---|
| M3 | 0 | 108.181 | 16.549 | 254.236 | 201.540 |
| M5 | 108.181 | 0 | 95.144 | 162.185 | 50.700 |
| M9 | 16.549 | 95.144 | 0 | 210.890 | 155.709 |
| WT | 254.236 | 162.185 | 210.890 | 0 | 171.747 |
| Neg | 201.540 | 50.700 | 155.709 | 171.747 | 0 |

Figure 2.5. LDA of with-in normalized from 30 aptamer set, 91.21% variance captured and Mahalanobis distance.

"within-slide" and "between-slide" normalization is performed, we observe a confusion matrix score of 100%. It is also noted that the Mahalanobis distance between M9 and M3 is lowest for both the unnormalized and within-slide only normalized, 20.2 and 16.6 as opposed to 24.1. As a result both within-slide and between-slide normalization was used.

# 4. Results

## 4.1 Initial aptamer analysis with 96 aptamers

Ninety-six of the anti-RT aptamers previously selected by Dr. Angel Syrett were screened for binding to the WT and the M3 enzymes. On a single slide there were 16 wells in which binding reactions could take place (Figure 2.6). Eight of the reaction

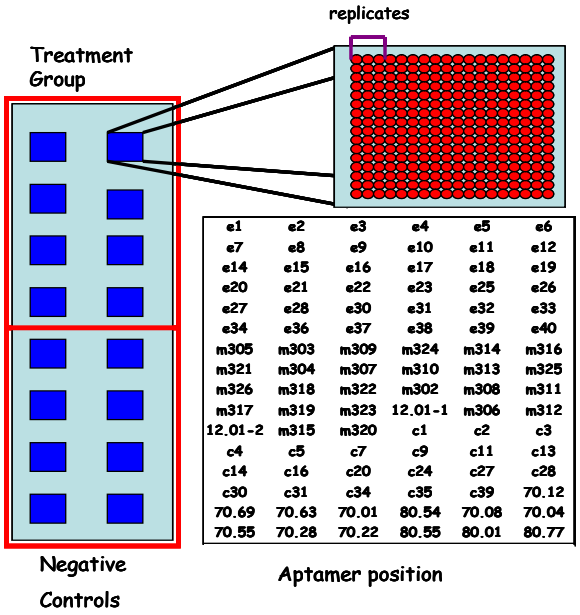



Figure 2.6. Layout of slides and reaction wells. Each name represents a triplicate printed on the slide.

wells, called treatment groups, were treated with either 850 pM wild-type RT or a mutant variant (M3, M5 or M9), while eight were treated with RT buffer only as a negative control. In total, there were 4 slides, one for each type of protein. Preliminary results indicated that at least some of the selected aptamers appeared to be semi-specific (Figure 2.7), we hypothesized that the arrays might allow recognition of the novel analytes M5 and M9 based on pattern recognition. It was found that a subset of the aptamers which

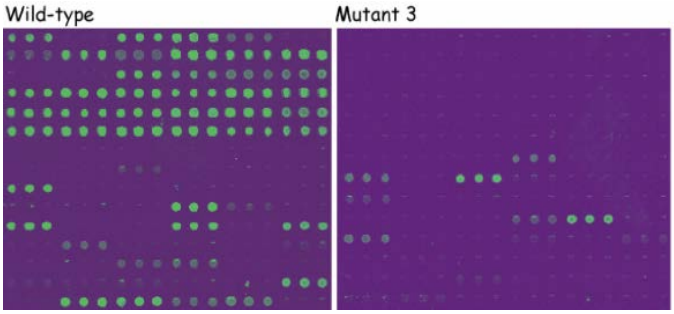



Figure 2.7. Cy3 channel of scanned image for wild-type RT and RT mutant M3, showing differential binding to aptamers. Courtesy of Angel Syrett.

were selected for specific binding to wild-type HIV-1 reverse transcriptase and the mutant variant M3 did indeed bind to mutant variants M5 and M9 (Figure 2.8). This is significant in that these aptamers had never previously been exposed to these variants.

The images presented in Figures 2.7 and 2.8 represent one of eight wells on a single slide treated with a particular reverse transcriptase variant. It was important to determine if all eight replicates in a single slide showed similar responses to the reverse

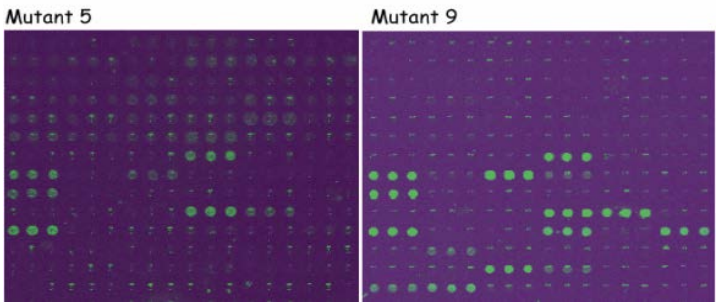
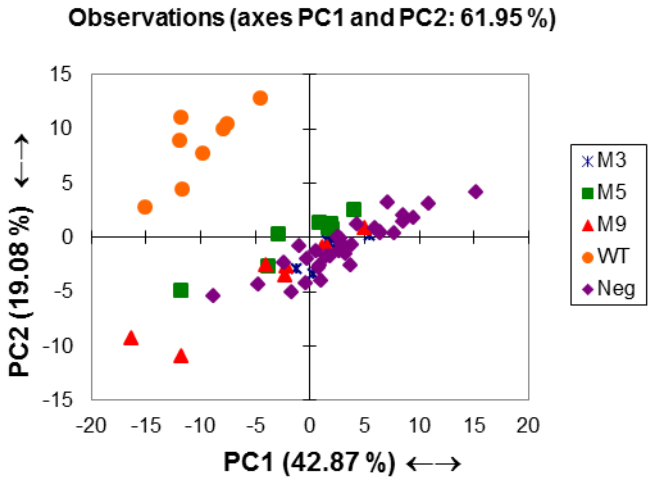
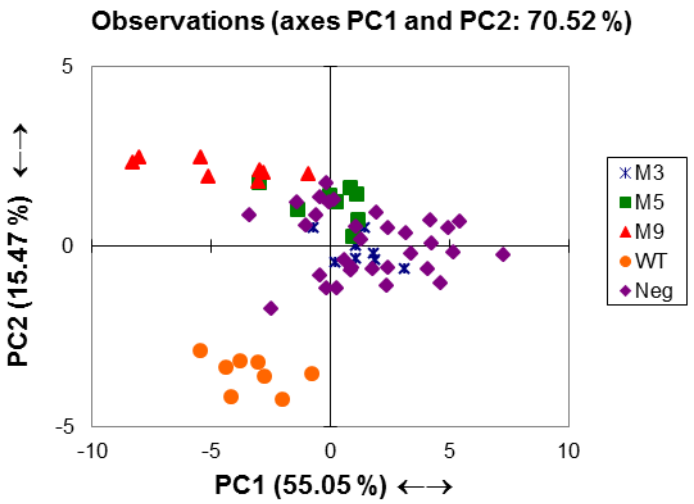



Figure 2.8. Cy3 channel of scanned image for RT mutants M5 andM9, showing differential binding to aptamers. Courtesy of Angel Syrett.




| | M3 | M5 | M9 | WT | Neg |
|---|---|---|---|---|---|
| M3 | 0 | 0.574 | 2.163 | 2.347 | 0.494 |
| M5 | 0.574 | 0 | 1.758 | 1.925 | 1.018 |
| M9 | 2.163 | 1.758 | 0 | 2.814 | 2.656 |
| WT | 2.347 | 1.925 | 2.814 | 0 | 2.538 |
| Neg | 0.494 | 1.018 | 2.656 | 2.538 | 0 |

Figure 2.9. PCA of within- and between-slide normalized, 61.93% variance captured, and Mahalanobis distance.




| | M3 | M5 | M9 | WT | Neg |
|---|---|---|---|---|---|
| M3 | 0 | 0.778 | 2.414 | 2.199 | 0.173 |
| M5 | 0.778 | 0 | 1.742 | 2.423 | 0.786 |
| M9 | 2.414 | 1.742 | 0 | 2.719 | 2.482 |
| WT | 2.199 | 2.423 | 2.719 | 0 | 2.371 |
| Neg | 0.173 | 0.786 | 2.482 | 2.371 | 0 |

Figure 2.10. PCA of 15 selected aptamers, 70.52% variance captured, and Mahalanobis distance.

transcriptases. Initial analysis of the data was unable to effectively discriminate the different RT variants using either the entire set of 96 aptamers (Figure 2.9). The fourth quadrant contains the entire set of wild-type arrays. However, the mutant arrays and

negative control arrays are spread over the first and second components, and there was significant overlap. In an attempt to improve the quality of discrimination, a subset of 15 aptamers predicted to be best at discriminating between proteins were used. The 15 aptamers were selected via a direct comparison method as described in Kitamura et al. (2009). However, like the full set of 96 aptamers, effective separation was not observed. Figure 2.10 shows that the discrimination was only moderately improved by using the selected aptamers. In the truncated set, both the wild-type and M9 points move away from the central cluster while M3, M5, and the negative control remain in the center. The Mahalanobis distance is a metric that measures the dissimilarity between clusters of data, and was used to give us a quantitative measurement of how similar or dissimilar each of the groups are**.** The percent of variance captured by the first two components increased from 61.95% to 70.52%. The inability to effectively discriminate the mutants may have been due to noise arising from non-specific, charge-based interactions with immobilized aptamers, irreproducible preparations of the large numbers of aptamers and slides, or an insufficient number of replicates.

### 4.2 AZT study

Azidothymidine (AZT) is a drug commonly used to treat HIV infection. It is an analog of thymidine and has a very strong affinity for HIV reverse transcriptase . When AZT is incorporated into an elongating DNA strand, it effectively terminates reverse

transcription (Meyer et al., 1999). Furthermore, when a nucleotide docks into the correct position of the reverse transcriptase and small conformational shift is induced

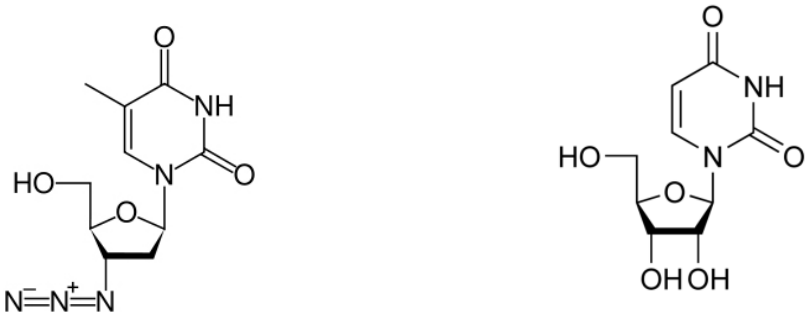



Figure 2.11. Structures of azidothymidine and thymidine.

(Spence et al., 1995). This leads to the question of whether or not an aptamer array can detect the presence of thymadine or AZT. Potentially the small conformational changes induced by the docking of AZT or thymidine into the active site of the reverse transcriptase could affect the way the reverse transcriptase interacts with some of the aptamers, thus generating unique patterns. To test this, 850 nmols of AZT or thymidine was incubated with wild-type HIV-1 reverse transcriptase. Each slide was set up such that there were eight wells which contained the enzyme solution, 4 wells as positive controls, and 4 wells as negative controls. There were two slides for each condition. The slides were treated and prepared as described previously. The enzyme mixtures were heated for 30 minutes at 37 °C in order to facilitate docking of the nucleotides in the active site of the enzyme. The enzyme alone was also incubated at 37 °C for 30 minutes to act as a control to determine if heating alone changed the binding profile. Figure 2.12 shows the

Cy3 binding patterns for enzyme with AZT, enzyme with thymadine and enzyme alone. It is not clear that there is any discrimination between the three samples, but it is possible

Figure 2.12. Array of 96 aptamers with only WT reverse transcriptase, with 850 nmols of AZT and 850 nmols of thymidine.

that the variations were too small to be distinguished. Figure 2.13 shows a PCA plot of the results of this experiment. There does appear to be some grouping relative to treatment group; the AZT treated enzymes predominantly occupy the lower right quadrant of the plot while the positive controls, enzyme-only, and thymidine-treated groups occupy the top half of the plot. The negative controls occupy the bottom left quadrant and the heated enzyme occupies the top right quadrant. Unfortunately there is a large amount of scatter present in this data, indicating that there is very little difference in each of the treatment groups.

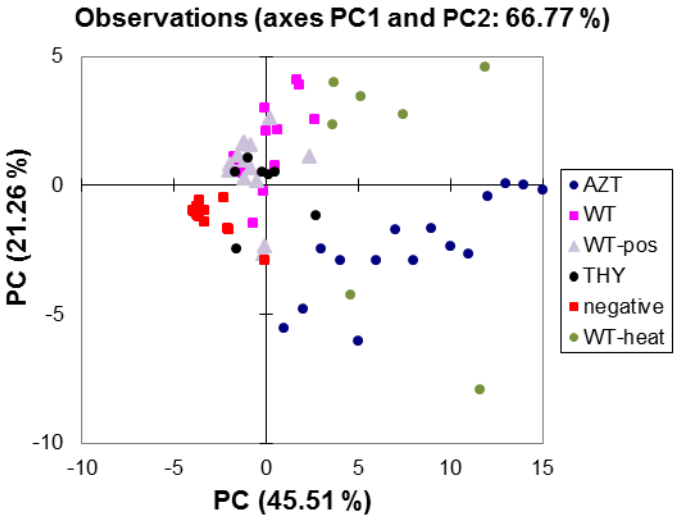



Figure 2.13. PCA plot of wild-type reverse transcriptase under various conditions: unheated RT, heated RT, 850 nmols AZT, 850 nmols thymidine, and a negative control.

In order to attempt to better discriminate the data, LDA was performed, and the results are found in Figure 2.14. In this case the scatter of the data has been significantly improved. Again, there does seem to be some grouping based on treatment group. While all the enzymes are in a single cluster the thymidine group is toward the bottom, the AZT group is toward the right, the heated enzyme is in the middle, and the unheated/positive controls are at the left. The grouping of all of the enzymes, along with the poor performance of the PCA, indicates that while it may be possible to detect the presence of AZT or thymidine with an aptamer array, substantial optimization would need to be done. It was decided not to pursue this route of research.

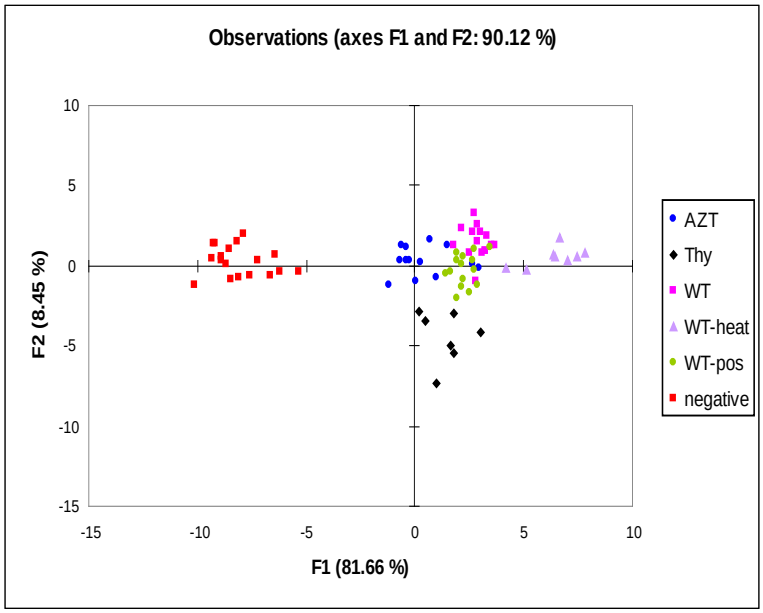



Figure 2.14. LDA plot of wild-type reverse transcriptase under various conditions: unheated RT, heated RT, 850 nmols AZT, 850 nmols thymidine and a negative control.

### 4.3 30-aptamer study

In order to more quickly generate more reproducible data, a more limited set of 30 aptamers was chosen for immobilization on new slides. Fifteen of the aptamers were selected because they were predicted to be the most useful in discriminating different protein variants. These aptamers in many instances were found to still be able to bind to several of the variants at once. Fifteen additional aptamers were included that appeared to be sensitive to the addition of AZT to RT, and thus that might also be sensitive to conformational changes in either the wild-type or mutant enzymes (Rekharsky et al., 2002).

The thirty selected aptamers were printed in 16 discrete locations on the slide; this formed 16 reaction wells that were independently probed with protein mixtures. Each

aptamer within each well was printed in sextuplet in order to better estimate statistical deviations in protein binding. Of the 16 wells, 8 were treated with either 850 pM WT RT or one of the three mutant variants (Table 2.1), four were treated with 850 pM RT as a positive control, and four were treated with buffer only as a negative control (Figure 2.15). In total, there were eight slides, two for each type of protein.

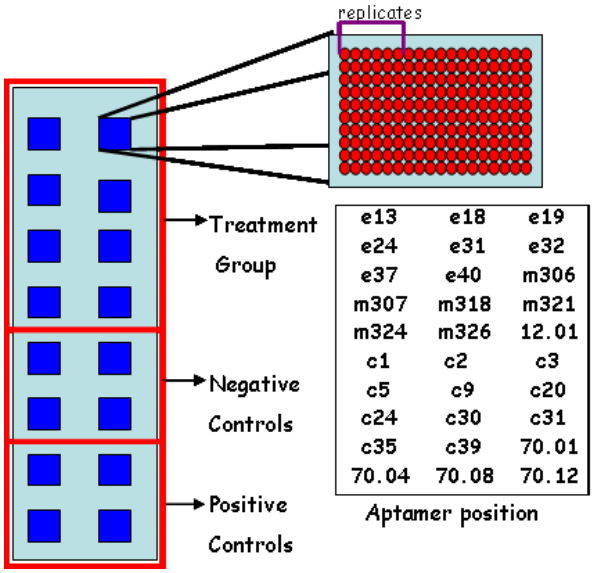



Figure 2.15. Slide treatment. Each small square represents a single reaction well, 16 per slide. Each well was identical and consisted of 30 aptamers printed in replicates of 6. The aptamer's position is a representation of where each aptamer was positioned relative to the others; each name represents a set of six replicates. Each slide was separated into three groups of wells. The top eight wells correspond to the treatment group; where one of the four HIV-RT variants was applied. The next four wells correspond to negative controls and the final four wells correspond to the positive controls.

Visual observation showed differential binding between the mutants, as seen in Figure 2.16. Spots which appear red indicate there is little protein captured by the aptamer, as only the Cy3 bound to the aptamer is visible. Spots which appear green indicate a large amount of protein capture as the Cy5 labeled secondary antibody obscures the Cy3 labeled aptamer. Black spots indicate little to no aptamer was deposited on the slide. All remaining spots have captured some amount of protein. As previously indicated, it was important to ensure that each of the eight replicates was measurably different from the others. Because of the highly multivariate nature of the data and noise

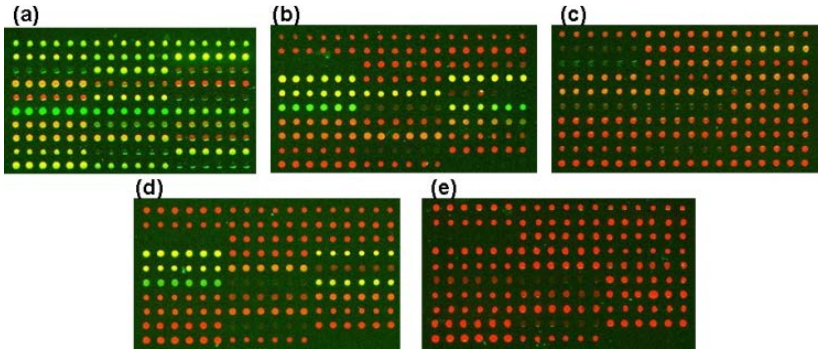



Figure 2.16. GenePix scan of the 30 aptamer set. Red corresponds to the Cy5 channel and the signal intensity is proportional to the amount of aptamer bound to the slide. "Black" spots indicated locations where little or no aptamer was deposited. If the background intensity exceeded the foreground intensity in either channel, the spots were excluded. Green corresponds to the Cy3 channel where the signal intensity is proportional to the amount of protein bound to the aptamer. a) Wild-type b) M3 c) M5 d) M9 e) negative control.

inherent to microarray experiments, a supervised learning algorithm, linear discriminant analysis (LDA), was selected for the data analysis. Figure 2.17 shows the results of the LDA on the within-slide normalized data. There are four distinct clusters of data. One cluster corresponds to the wild-type protein, which is separated from the mutant proteins and negative controls across the first and second components. The buffer-only negative control and M5 mutants are separated from the WT, as well as M3 and M9 proteins, primarily across the second component. However, there is minimal separation of the

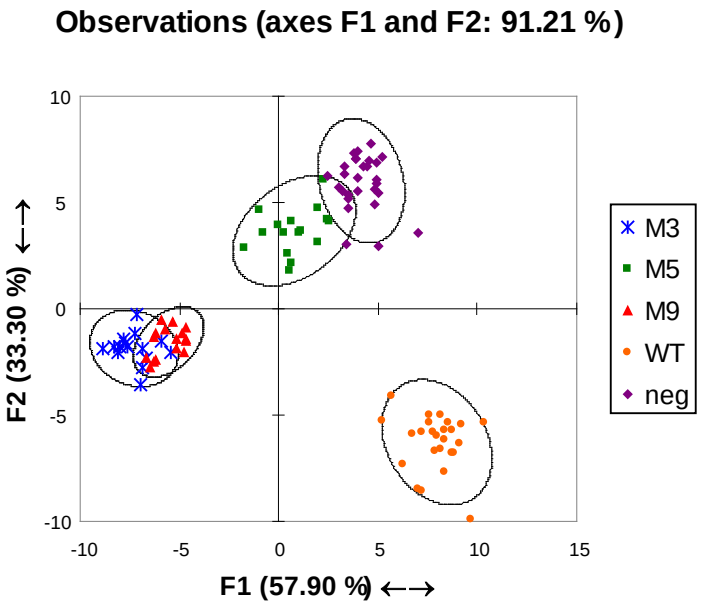



Figure 2.17. LDA plot of normalized 30-aptamer dataset. Ellipses represent 95% confidence intervals.

negative control from M5, and of M3 from M9. As might be expected based on these analyses, the visual pattern of binding for M5 is similar to that of the negative control,

and the visual patterns for M3 and M9 are similar (Figure 2.16). By examining the factor scores for the first, second, and third components it became apparent that clusters are separated across the third component as well as the first two (Figure 2.18). By including a third component, one can see that mutant M5 was primarily separated from the others across the third component.

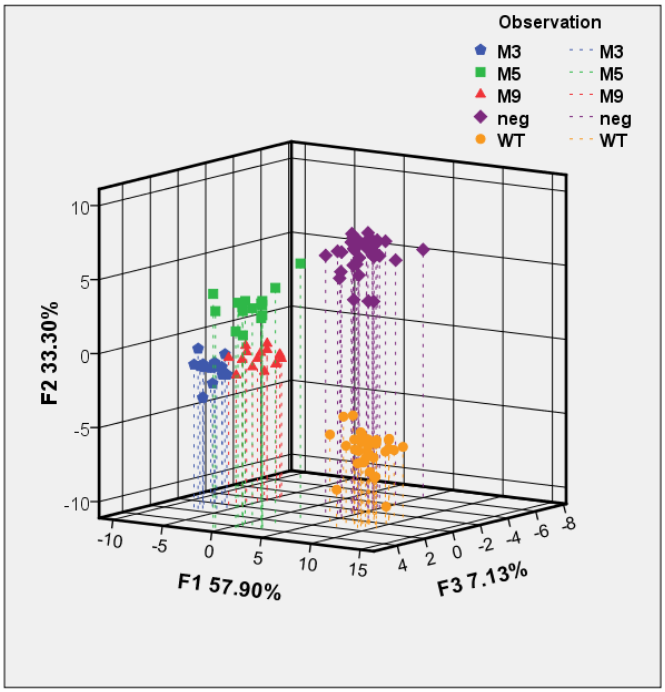



Figure 2.18. LDA plot of normalized 30-aptamer data set including the third component.

| from \ to | M3 | M5 | M9 | WT | Neg | Total | % correct |
|---|---|---|---|---|---|---|---|
| M3 | 15.98667 | 0 | 5.813333 | 0 | 0 | 21.8 | 73.33% |
| M5 | 0 | 18.68571 | 1.557143 | 0 | 1.557143 | 21.8 | 85.71% |
| M9 | 0 | 0 | 21.8 | 0 | 0 | 21.8 | 100.00% |
| WT | 0 | 0 | 0 | 21.8 | 0 | 21.8 | 100.00% |
| neg | 0 | 0 | 0 | 0 | 21.8 | 21.8 | 100.00% |
| Total | 15.98667 | 18.68571 | 29.17048 | 21.8 | 23.35714 | 109 | 91.81% |

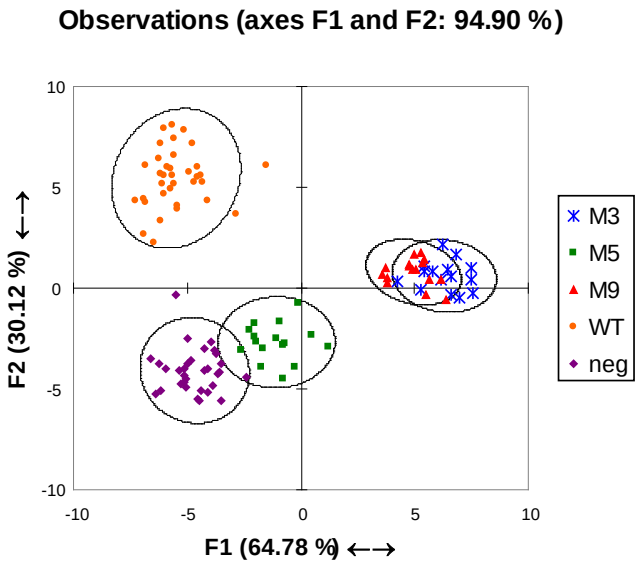
Table 2.3 Cross validation results.

Furthermore, there was also modest improvement in the visual discrimination between M9 and M3. In order to estimate the extent to which the model was predicting itself, a leave-one-out cross-validation test was performed (Table 2.3**).** In this analysis each sample was removed from analysis and the remaining points were used to build the model. The model then attempted to predict which grouping the omitted point belonged to. M9, wild-type, and the negative control were correctly predicted 100% of the time. However 7.15% of the time a M5 well was incorrectly predicted to be M9 and 7.15% of the time it was predicted to be a negative control. This resulted in an overall accurate prediction for M5 85.71% of the time. Interestingly, while M9 was never incorrectly predicted to be M3, M3 was mistaken for M9 26.67% of the time. This implies that there

is insufficient difference between the patterns generated by M3 and M9 to effectively separate the two.

### 4.4 Truncated aptamer set

Previously, modest improvement was observed when only a subset of the aptamers printed on a slide was used to build a model. It was hypothesized that improvement could be achieved by using only the aptamers which were found to be best for discriminating the mutant variants; selected as previously described. In this situation, the resolving power would be expected to improve by limiting extraneous signals from irrelevant aptamers. The results are shown in Figure 2.19. As with the full set of aptamers, there exist four separate data clusters. However, M3 and M9 remain so close as to be indistinguishable, though M5 and the negative control have separated somewhat. Wild-type remains well separated from the mutant variants. When the third component is considered, very little improvement is observed (Figure 2.20). This is to be expected as the first two components account for 94.90% of information used to segregate the groups. This leads to the conclusion that all 30 aptamers are indeed required for optimal discrimination with this method.




| | M3 | M5 | M9 | WT | neg |
|---|---|---|---|---|---|
| M3 | 0 | 72.921 | 6.472 | 163.179 | 144.487 |
| M5 | 72.921 | 0 | 55.482 | 89.621 | 27.754 |
| M9 | 6.472 | 55.482 | 0 | 131.939 | 119.262 |
| WT | 163.179 | 89.621 | 131.939 | 0 | 93.740 |
| neg | 144.487 | 27.754 | 119.262 | 93.740 | 0 |

Figure 2.19. LDA of 15 selected aptamers and results of leave one out cross validation.

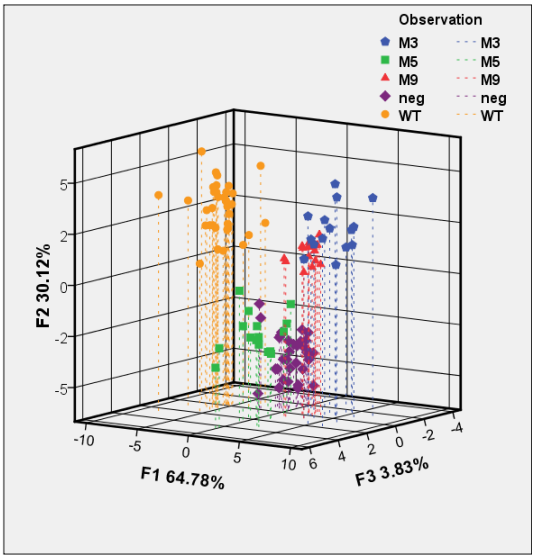



Figure 2.20. LDA plot of 15 selected aptamers with third component included.

## 5. Discussion

Taken together, these results support the hypothesis that aptamers can be used as semi-specific receptors for identifying protein variants. The idea that an aptamer microarray can detect proteins is not a new notion, nor is the use of semi-specific receptors a new strategy. However, the utilization of aptamers as semi-specific receptors on a microarray is a new application. Perhaps the most significant revelation from this work is the ability of the array to effectively discriminate proteins to which they had not previously been exposed to. This indicates that an array of aptamers could potentially be used to create unique fingerprints for a wide variety of targets. However, it is not

immediately obvious what parameters should be used for the construction of arrays that provide broad recognition. For example, even though aptamers that recognized M3 were also somewhat related to aptamers that recognized M9, these proteins do not share any amino acid substitutions. It is unclear what characteristics are driving the affinity in this case. It is possible that aptamers that bound both of these proteins recognized conformational changes induced by the mutations rather that the mutations themselves. This notion is supported by the work done to identify enzymes that were associated with a nucleotide. In this situation, the only possible variation in the enzymes used was conformational change. Despite only limited discrimination being achieved, there does seem to be some support that conformation is important to the affinity between enzyme and aptamer.

Finally, limiting the array to those 15 aptamers predicted to provide the best discrimination of the mutant variants did not improve discrimination. It is probable that at least some of the aptamers selected for mutant discrimination were specifically affected by the amino acids substitutions, while some were more driven by conformational changes. However, 15 aptamers which may or may not have been sensitive to conformational changes did not prove to be sufficient in this experiment. Rather, including those aptamers thought to be most influenced by shape were also required for success.

Despite the promising results from this study, it was decided not to pursue aptamer microarrays as a semi-specific array platform. Significant variation between experimental replicates severely limited the power of this assay. Furthermore, the

complex system devised for detection - anchor, aptamer, probe, enzyme, primary antibody, secondary antibody - proved to be cumbersome and required a significant amount of work (a large portion of two graduate students' work) to achieve success.

Despite the difficulties this work encountered, it can still be considered a success in the advancement of aptamers as semi-specific receptors. We have shown that a panel of aptamers immobilized in a microarray format can be used to discriminate between similar proteins based on chemometric methods. Furthermore, these aptamers could distinguish proteins that differed by as few as 4 amino acid substitutions, even when the aptamers were not specifically selected against a given target protein. Because aptamers combine complex recognition features with synthetic tractability, they may prove to be particularly useful for the production of cross-reactive arrays for biomedical testing, bio-defense applications, and food quality monitoring (Peng et al., 2010; Phillips et al., 2008; Taitt et al., 2002; Zhang et al., 2007).

The problem with this work applies only to the microarray platform, but not to the performance of the aptamers themselves. Eliminating the need for immobilization strategies and visual signals for aptamer target interactions could significantly simplify the use of aptamer panels for biomolecular discrimination. The elimination of visual signals could further advance the use of aptamers in these sorts or routines. Inherently visual signals are limited by the ability to interrogate the signal itself. Certainly there are methods where fluorescence is measured across many wavelengths to create a unique pattern (Kitamura et al., 2009), but these methods are still limited by the optical

resolution of the instrument. In the following chapter, we will explore using aptamer sequencing as method of signal interrogation for complex target discrimination.

## 6. Author contributions

SS Devised the experiments, prepared the aptamers, prepared the microarrays, performed data analysis, and wrote the manuscript. AS selected aptamers and assisted in experimental design. AP and SB prepared mutant proteins. AE and EA oversaw experiments and provided scientific direction.

# Chapter 3: Exploring the use of Aptamers as Non-specific Biomolecular Receptors

*This chapter is being adapted for a manuscript to be submitted to Nature Chemical Biology*

## 1 Introduction

Literature often touts aptamers as being highly specific, alternatives to antibodies (Ellington and Szostack, 1990; Tuerk and Gold, 1990; Jayasena, 1999; Proske et al., 2005; Strehlitz et al., 2008; Long et al., 2008; Xiao et al., 2008). Additionally, aptamers have the unique ability to be rapidly selected for a wide range of targets, many of which are not compatible with antibodies. Certainly, there are a number of aptamers that do display antibody-like behavior and perform quite well under any number of different conditions (Geiger et al., 1996; Jellinek et al., 1993, Osborne et al., 1997). However, as the number of aptamers developed has expanded, it is becoming apparent that aptamers will not unseat antibodies as the champions of sensitivity and selectivity. This is evidenced by there being only one FDA approved therapeutic aptamer (Ni et al., 2011), while there are dozens of FDA approved antibody based therapeutics. Like antibodies, aptamers are subject to cross reactivity, which limits their utility in very complex environments. Additionally, because of limited chemical interactions, an oligonucleotide can carry out as opposed to an antibody, antibodies outperform aptamers when it comes to affinity (Keefe et al., 2010). This is why so few aptamers have found a future outside of the research lab- for most applications, antibodies simply perform better.

However, the limited chemical interactions carried out by aptamers can be used as an advantage. In complex solution sensing routines, which use highly specific receptors, one can only detect and discriminate as many compounds as there are receptors. The number of unique binding patterns which can be generated (when concentration data is not considered) from such a routine is defined by the equation $2^n$. While technically this formula is mathematically limited only by the number of receptors available, it would be impossible to generate a receptor for all possible targets in a complex solution. Rather, patterning all compounds in a complex solution can be more readily achieved with cross reactive receptors (Albert et al., 2000; Anslyn, 2007). This type of routine is a more realistic approach to detection and discrimination of complex targets as it does not rely on highly specific receptors, which can be difficult the generate on a large scale. Instead, each receptor binds differentially to each possible target in the complex sample. This allows for the generation of a unique pattern, or fingerprint, for each different sample. Aptamers, by virtue of their simple chemistries, make ideal candidates for cross-reactive receptors. However, to date, there has been very little work in developing aptamers for cross-reactive sensing routines.

There are currently a number of cross-reactive platforms, dedicated to discriminating a wide variety of different targets, including the use tin oxide ($SnO_2$) sensors to detect toxic gasses, aptamer microarrays to discriminate protein variants, hydrophobic dyes to detect organic compounds in water, and functionalized gold nanoparticles for cell discrimination (Mishra and Agarwal, 1998; Bajaj et al., 2009; Zhang and Suslick, 2005; Stewart et al., 2011). With-in the field of cross reactive

sensors, there are a number of platforms dedicated to discriminating biological targets. This is driven by thriving research into biomarkers. Biomarkers for disease states have become commonplace in today's modern medicine. A patient can go to their physician, and with a quick blood draw find out their levels of medically significant molecules such as c-reactive protein, prostate specific antigen or interleukin-1b (Yeh and Willerson, 2003; Barry, 2001; Li et al., 2004). Many of these markers have proven to be invaluable to the detection and discrimination of numerous chronic diseases. Recently the number of biomarkers available for disease state detection and discrimination has increased substantially. Unfortunately methods for discriminated disease states with biomarkers have not kept up.

Approaches to this problem historically have been dominated by the use of immunoassays , which utilize one or more antibodies to act as receptors for the molecule of interest. Perhaps the most important immunoassay is the ELISA or enzyme-linked immunosorbent assay. In this approach antibodies are immobilized in a well plate and incubated with the fluid of interest. However, traditional ELISAs in well plates, while powerful, are limited by the sample volumes needed and the number of targets that can be interrogated simultaneously. Antibody microarrays, where antibodies are immobilized on functionalized glass slides, use much smaller sample volumes and allow for significantly more targets to be analyzed in a single experiment; high density antibody microarrays, for example, have been shown to detect tens of thousands of targets (LeBaron et al.,2005).

Each of these approaches, however, are limited by the detection strategy. Frequently, for both ELISA and microarray experiments, antibodies are paired together such that one antibody captures the target molecule and another antibody, carrying a fluorescent molecule, binds the target to allow for detection. Substantial screening and time must be dedicated to creating two non-competitive antibodies for this type of application. Alternately, the target itself may be modified to carry a fluorescent signal, however there is no guarantee the antibody receptor will still bind the modified target. Perhaps the most significant issue with an assay reliant on fluorogenic signals is the limited color space they are subject to. There are only a limited number of dyes which generate sufficiently unique emissions to be used concurrently. Furthermore, current instrumentation can detect only a limited number of emissions concurrently (Waggoner, 2006). Thus, the logistics of generating antibodies in addition, to limitations in color space, negatively impact the utility of immunoassays for highly multiplexed detection schemes.

An approach to biomolecular detection that does not directly involve the use of a fluorogenic, signal has been to amplify a nucleic acid sequence somehow associated with the target. The bio-bar-code method uses DNA functionalized gold nanoparticles to target sequences of interest. The gold nanoparticles carry a unique identifier oligonucleotide (a bar code) that can be amplified and detected via a microarray or amplification (Nam et al., 2004). The bio-bar-code method was extended to aptamers in order to alleviate the need for gold nanoparticles. Instead, the aptamer itself binds to the target and its presence can be detected through amplification (Lau et al., 2010). Sequencing is also a method

which has been used to detect the presence aptamers as an alternative to fluorogenic signals. Turner et al. used next generation sequencing (NGS) to measure the concentrations of four different proteins using four specific aptamers (Turner et al., 2010). All of the previous approaches share one major limitation; none of them are truly compatible with a highly complex system. In each case a specific molecule of some sort was dedicated to a specific target. As previously discussed, this approach is simply too technically complicated to be realistic. Instead, we propose the use of cross-reactive aptamers for the discrimination of complex targets. The complex composition of the cell surface and the availability of validated cell surface aptamers made cells an ideal choice for a model complex target.

Theses aptamers will constitute a panel of cross-reactive sensors for cellular discrimination. In this panel each aptamer will be present as a discrete proportion of the whole. After the panel has been allowed to bind to the cell surface and non-binders removed, what will remain is a unique distribution of aptamers whose abundance is dictated but each aptamer's affinity of the surface. Next generation sequencing (NGS) technology allows for the sequencing of an enormous number of sequences simultaneously. By amplifying and sequencing the aptamers collected from the assay, a characteristic distribution of aptamers can be used to discriminate the targets.

In this study we outline a method for discriminating four different cancerous and normal cell types. It shows that using NGS technology to interrogate a panel of 46 cross-reactive aptamers gleaned from the literature (Madsen et al., 2010; Chauveau et al., 2007; Cerchia et al. 2009; Chu 2006; Lee 2007; Li 2011; Hicke et al., 2001; Dollins et al., 2008;

Lupold et al., 2002; Daniels et al., 2002; Jeong et al., 2001; Cerchia et al., 2005; Davis et al., 1998; Kraus et al., 1998; Mi et al., 2005; McNamara et al., 2008; Barfod et al. 2009; Biesecher et al., 2005; Liu 2009; Magalhães et al., 2012) is indeed an effective approach to complex target discrimination. Theoretically this panel of aptamers, paired with the immense data collection power of NGS, could effectively generate a unique pattern for any cell type; thus eliminating the need for unique assays for each cellular target.

# 2 Materials and Methods

## 2.1 Aptamer selection

2'-fluoro modified RNA oligonucleotides have proven to be exceptionally resistant to nuclease activity (Kawasaki et al., 1993). Furthermore, mutant T7 transcriptases have been developed capable of incorporating the modified nucleotides in to a new RNA strand form a standard DNA template. This makes this particular type on unnatural nucleic acid particularly well suited for use as aptamers.

There are hundreds if not thousands of published sequences for 2'-F RNA aptamers targeting a huge variety of different targets. In order to explore the utility of an aptamer panel for complex target discrimination, cancer cells were selected to represent the complex target. Many 2'-F aptamers exist for targets that could be expected to be express on a cell surface and the panel used for this study draws on that knowledge base.

There were three broad categories of nuclease resistant aptamers used in this study: aptamers selected to directly bind cells, aptamers selected to bind molecular targets found on the cell surface, and aptamers not expected to bind cells (Table 3.1).

The first category included aptamers selected to targeted to the U87MG glioma line, PC3 prostate cancer line, H358 non-small cell lung carcinoma (NSCLC) line, and H562 a small cell lung carcinoma (SNLC) line. Many of these aptamers were selected to show broad target ranges. For example the aptamer "GL44" was originally selected against U87MG cells, but also bound to the U87MG, LN-18, LN-229, U87MGΔEGFR, and TB10 cell lines.

The second category included aptamers selected against targets that are generally found at the cell surface, including the extracellular matrix proteins, TN-C and PAI-1; receptors such as the RET kinase, EGFR, EGFRΔIII, OX40, VCAM-1 CD4, NTS-1, PfEMP-1, and αvβ3; the glycoproteins PSMA, and 4-1BB and the carbohydrate Sialy lewis X. While not all of the selection targets were human proteins, cross-binding to human proteins had been demonstrated in many instances.

Finally two aptamers were included as negative controls. Aptamers "nt" and "G31" were known not to bind cells and were scrambled versions of the NTS-1 and αvβ3 aptamers.

| Type | Target | Name | Working name | Citation |
|---|---|---|---|---|
| Cell | U87MG | GL17 | C9 | Cerchia et al., 2009 |
| Cell | U87MG | GL44 | C10 | Cerchia et al., 2009 |
| Cell | U87MG | GL56 | C11 | Cerchia et al., 2009 |
| Cell | PC3 | PC301 | C12 | Chu, 2006 |
| Cell | PC3 | PC304 | D1 | Chu, 2006 |
| Cell | Clone 2 | H526 | D2 | Lee, 2007 |
| Cell | Clone 4 | H526 | D6 | Lee, 2007 |
| Cell | Clone 7 | H526 | D5 | Lee, 2007 |
| Cell | Clone 9 | H526 | D3 | Lee, 2007 |
| Cell | Clone 10 | H526 | D4 | Lee, 2007 |
| Cell | A6 | H358 | D7 | Lee, 2007 |
| Cell | E5 | H358 | D9 | Lee, 2007 |
| Cell | E7 | H358 | D8 | Lee, 2007 |
| Cell | D4 | PC12/MEN2B | E12 | Cerchia et al., 2005 |
| Cell | D24 | PC12/MEN2B | F1 | Cerchia et al., 2005 |
| Cell | E9P2-1 | U251 | D12 | Hicke et al. 2001 |
| Cell-Surface Protein | 12.11 | VCAM-1 fragment | C7 | Chauveau et al., 2008 |
| Cell-Surface Protein | 12.23 | VCAM-1 fragment | C8 | Chauveau et al., 2008 |
| Cell-Surface Protein | E07 | EGFR | D10 | Li et al., 2011 |
| Cell-Surface Protein | E-9 | TN-c | D11 | Hicke et al. 2001 |
| Cell-Surface Protein | TN-9 | TN-C | E1 | Hicke et al., 2001 |

Table 3.1 List of aptamers and their targets.

| | | | | |
|---|---|---|---|---|
| Cell-Surface Protein | 9.4 | OX40 | E3 | Dollins et al., 2008 |
| Cell-Surface Protein | 9.8 | OX40 | E2 | Dollins et al., 2008 |
| Cell-Surface Protein | A9 | recombinant PSMA | E4 | Lupold et al., 2002 |
| Cell-Surface Protein | A10 | recombinant PSMA | E5 | Lupold et al., 2002 |
| Cell-Surface Protein | P19 | NTS-1 | E6 | Daniels et al., 2002 |
| Cell-Surface Protein | P112 | NTS-1 | E7 | Daniels et al., 2002 |
| Non-binder | G31 | Scramble | E8 | Daniels et al., 2002 |
| Other | 5(9) | $SLe^X$ | E9 | Jeong et al., 2001 |
| Other | 2(2) | $SLe^X$ | E10 | Jeong et al., 2001 |
| Other | 5 | PAI-1 | C5 | Madsen et al., 2010 |
| Other | 40 | PAI-1 | C6 | Madsen et al., 2010 |
| Other | C1 | Ubiquitous internalizer | E11 | Magalhães et al., 2012 |
| Cell Surface protein | 9-I | CD4 | F2 | Davis et al., 1998 |
| Cell Surface protein | 12-II | CD4 | F3 | Davis et al., 1998 |
| Cell Surface protein | 8-I | CD4 | F6 | Krause et al., 1998 |
| Cell Surface protein | 33-III | CD4 | F7 | Krause et al., 1998 |
| Cell Surface protein | avB3 | αvβ3 | F8 | Mi et al., 2005 |
| Other | E05 | PfEMP-1 | F11 | Barfod et al., 2009 |
| Other | B2 | PfEMP-1 | F12 | Barfod et al., 2009 |
| Cell Surface Protein | M12-23 | 4-1BB | F10 | McNamara et al., 2008 |
| Cell Surface Protein | C6 | Complement C5 | G1 | Biesecker et al., 1999 |
| Cell Surface Protein | H2 | Complement C5 | G2 | Biesecker et al., 1999 |
| Cell Surface Protein | E27 | Unglycosylated EGFRvIII | G3 | Liu et al., 2009 |
| Cell Surface Protein | E17 | Unglycosylated EGFRvIII | G4 | Liu et al., 2009 |

Table 3.1, cont.

## 2.2 Aptamer generation

For this study, all aptamers were selected from the literature or obtained from members of the Ellington lab. Each aptamer template sequence was modified to remove the original T7 promoter region and add forward and reverse constant primer regions which were compatible with next generation sequencing platforms. Each template sequence was synthesized using the MerMade 192 synthesizer (Bioautomation, Plano, TX), and diluted to 100µM in TE buffer. One µl of each synthetic aptamer template was amplified though PCR in order to maintain a double stranded stock solution. One µg of each template was transcribed to generate the 2'-F modified RNA aptamer using the Dura-scribe reverse transcription kit (Epicenter, Madison, WI) and treated with DNAse to remove any residual template. The product of each transcription reaction was purified on an 8% SDS page gel with 7M urea. The 2'-F RNA was eluted from the acrylamide in water over night and precipitated using ethanol precipitation. Each aptamer was diluted to 10nM in binding buffer (PBS supplemented with 5mM $MgCl_2$). Just prior to each experiment, each aptamer was folded alone or with oligonucleotides complementary to the 5' and 3' constant regions by at 75ºC for 3 minutes, followed by a ramp down of 1ºC a second to 25ºC. After folding, each aptamer was mixed at equimolar proportion to create the aptamer panel.

## 2.3 Cell assay

Each cell culture was grown in Dulbeccos modified eagles medium (DMEM) (Life Technologies Grand Island , NY) in 6 or 12 well culture plates to ~80% confluence.

U87MGvIII cells carry a plasmid allowing for the expression mutant EGFRvIII. In order to maintain the plasmid, the cells were grown in the presence the antibiotic G418 at 400 µg/ml. The media was removed and each well was washed 3x with binding buffer. One (6 well plate) or two (12 well plate) wells on each plate were treated with trypsin and the resultant cells were counted to determine the approximate number of cells in each well. 4.6 µl (0.01pmols of each aptamer), 46 µl (0.1pmols), 460 µl (1pmol) or 920 µl (2pmols) of the mixed aptamer panel was mixed with binding buffer to a total volume of 1 ml. One ml of this solution was added to each experimental well for each $1x10^6$ cells. One to two wells on each plate were reserved as negative control wells and were incubated with 1ml of binding buffer per each $1x10^6$ cells. The cells were incubated at room temperature with agitation for 30 minutes. At the end of the incubation period, the aptamer solution was removed and each well was washed 3X with binding buffer. The cells were lysed and total nucleic acids were collected using the MasterPure total nucleic acid system (Epicenter, Madison, WI). Aptamers in 5µg of the total nucleic acid solution from the cells were reverse transcribed using the Superscript III first stand synthesis system (Life Technologies Grand Island, NY). For each cell line used, 1µl of the naïve panel was mixed with 5µg of total nucleic acid solution obtained from negative control wells. This solution was also reverse transcribed. From each reaction, 2µl were passed to PCR to regenerate dsDNA templates compatible for NGS. The dsDNA templates were separated from primers, which can lower NGS results, through excision from a 2% TAE agarose gel. The dsDNA was purified from the agarose using the Wizard DNA clean-up system (Promega, Madison, WI). Aliquots of each of the purified samples were submitted to the

University of Texas NGS facilities for sequencing on either the SOLiD (Life Technologies Grand Island, NY) platform or the Illumina HiSeq platform (Illumina, San Diego, CA).

### 2.4 FACs analysis

Fluorescence activated cell sorting (FACs) was used to identify the best positive control choices and to explore the behavior of individual aptamers alone and as a panel. One pmol of an aptamer was mixed with 1pmol of a biotin capture probe or capture probe and oligonucleotides complementary to the 3' constant regions and mixed 1:1 with binding buffer. The mixture was incubated at 75ºC for 3 minutes and ramped down 1ºC a second to 25ºC, to allow the capture probe to anneal the probe/oligonucleotides and allow the aptamer to refold. Equimolar streptavidin/phycoerythrin was added to the aptamer and incubated at room temperature for 3 minutes. Cells grown to ~80% confluence were trypsanized and washed in binding buffer; the cells were resuspended in 100µl of binding buffer. For each 100,000 cells, either 1pmol or 10pmol of the labeled aptamer was added. In experiments where the aptamer panel was also included, each unlabeled aptamer was folded independently from the panel and added to the cells separately from the remaining panel at a concentration of 1pmol per $1x10^6$ cells. 450µl (1pmol of each aptamer) per $1x10^6$ cells of unlabeled panel was then added to the cells. The cell/aptamer solution was allowed to incubate for 30 minutes at room temperature with agitation. After incubation the cells were pelleted and unbound aptamer was removed. The cells were

washed 2x with binding buffer and analyzed with the LSR Fortessa cell analyzer (BD Biosciences San Jose, CA).

### 2.5 Real time analysis

Real time analysis was used to identify the best positive control choice and verify sequencing and real time results. Applied Biosystems SYBR green RT-PCR kit was used for these experiments (Life Technology Grand Island, NY). Five µl of total nucleic acid solution (prepared as described in 3.2.2) was mixed with SYBR green master mix and primers were added to a final concentration of 300nM. In some case a standard reference curve, between 1amol and 1pmol, was included to allow for quantitation of an aptamer's abundance. The instrument used for analysis was the Viia7 Real-Time PCR system (Life Technologies, Grand Island, NY).

### 2.6 Analysis of NGS data

For the single cell line experiments the SOLiD sequencer platform was used, (Life Technologies, Grand Island, NY). For all remaining analyses the HiSeq sequencer platform was used (Illumina, San Diego, CA). For SOLiD applications a reference library containing all possible aptamers was generated and converted to colorspace. All reads were done using a paired end 50/35/10 (forward, reverse, barcode) protocol. For the HiSeq platform all reads were done using a paired end barcode + 2x100 (100 reads forward and reverse) and a native reference library was used. The reads were aligned and mapped using BWA to the reference library and the number of hits for each aptamer in

each sample was recorded. Sequences which did not align, aligned to multiple aptamers or where unexpectedly long were excluded from analysis. The abundance of each aptamer was expressed as a ratio of the number of hits for a single aptamer over the total number of hits for all aptamers in a single sample. For the PCA plot each aptamer was normalized to the naïve panel by taking the ratio of the aptamer abundance in an experimental sample over the abundance of the aptamer in the naïve panel. The data was mean centered and passed to XLstat for PCA analysis (Pearson). The fold change of each aptamer was calculated as the $\log_2$ ratio of the abundance of an aptamer in an experimental sample over the abundance of an aptamer in the naïve panel.

# 3 Results

## 3.1 Preliminary real time results

Since we anticipated that chemometric data from deep sequencing might be skewed by sequencing biases, we performed preliminary experiments with real-time PCR that could serve as a guide for interpreting the sequencing data. Two cell lines were chosen for this study, A431 and MDA-MB-435. A431 is an epidermoid carcinoma that is often used in studies because of its abnormally high levels of epidermal growth factor receptor (EGFR), (Ullrich et al., 1984) an ErbB receptor family member whose overexpression is frequently found to be a biomarker for tumors (Zhang et al., 2007). In contrast, MDA-MB-435 is a breast carcinoma line that is frequently paired with the A431 line as it demonstrates 4 times lower expression of EGFR (Kempiak et al., 2003). We anticipated that these cell lines might be distinguished using an anti-EGFR aptamer (D10,

Table 3.1) and real-time PCR analysis, since this aptamer had already demonstrated differential binding via FACS analyses (Li et al., 2011).

Three other aptmers were also chosen for initial real-time PCR analysis of differential binding: F7, E11, C7, and E12. Aptamer F7 binds to CD4, a glycoprotein that is predominantly found on the surface of leucocytes (Maddon et al., 1985). CD4 is a member of the Immunoglobulin super family (IgSF) which includes similar members express on many different cell lines. There is no evidence that CD4 is expressed on A431 or MDA-MB-435 cell lines, though other members of the superfamily, such as ICAM are (Teixeira, 1999). Aptamer E11 is known to bind to and be internalized by a wide variety of cell lines (Magalhães et al., 2012). Aptamer E12 was selected against the RET kinase expressed on the surface of PC12 cells. RET is a ubiquitous receptor that is known to be expressed on many different cell lines (Robinson 2000).

To carry out these experiments a panel of 46 aptamers at various concentrations, suspended in PBS supplemented with 5uM $MgCl_2$ (0, 0.01, 0.1, 1, or 2 pmols of each aptamer per $1x10^6$ cells) were incubated with adherent cells in six well plates at room temperature for 30 minutes. Unbound aptamer was washed away and the cell lysate containing the bound aptamers was collected. The recovered aptamers were reverse transcribed and then subjected to real-time PCR analysis. In addition 1nmol of a pure panel of the same aptamers was mixed with 5ug of untreated lysate and reverse transcribed and analyzed by real-time PCR as well. Each of the experimental aptamers was amplified in the real time analysis with primers specific to the random region of each aptamer. This allowed the amplification of a single aptamer out of the panel. The

abundance of each aptamer isolated from treated cells was measured and well as the abundance of each aptamer amplified from the naïve panel suspended in untreated lysate was measured. The ratio between the abundance from the treated cells over the naïve panel was used to determine differential binding. Those values with a higher ratio value were indicative of better aptamer affinity, while low ratio values indicated low aptamer affinity for the cell line

As expected, D10 showed enrichment on the A431 cell line and depletion on the MDA-MB-435 line (Figure 3.1). This is seen as a fold increase over panel (positive values) for A431 and fold decrease below panel (negative values) for MDA-MB-435. The error bars represent cumulative standard deviation from the triplicate technical replicates and the duplicate experimental replicates. Like D10, aptamer F7 showed some enrichment of the aptamer on the A431 line. Conversely aptamers E11 and E12 (seemed to show preference for the MDA-MB-435 line. Overall, these data revealed expected differences between aptamer binding to cellular targets, and provided a touchstone for the eventual interpretation of deep sequencing analyses.

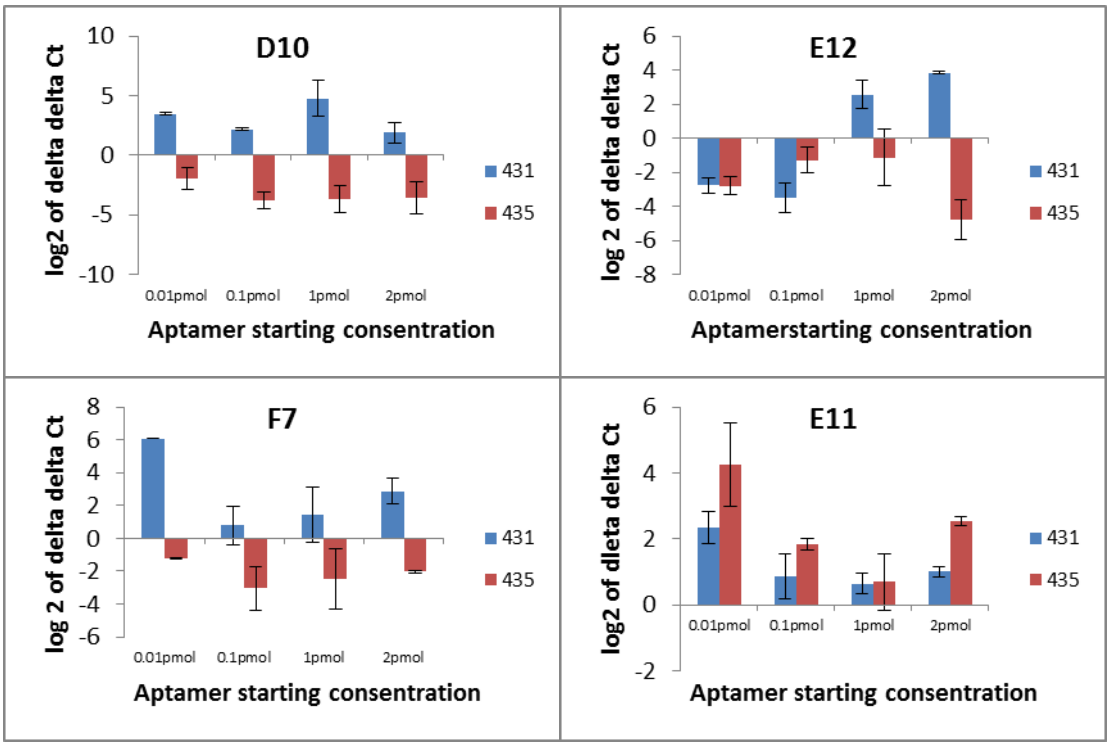



Figure 3.1 Preliminary real-time results for 4 selected aptamers. Aptamers D10, E12 and F7 show enrichment for A431 and depletion for MDA-MB-435. D11 show enrichment for both. Error bars represent standard error.

### 3.2 Analysis of a single cell line, A431

Forty-six 2'-F modified RNA aptamers, that target either cells or proteins expected to be expressed on the surface of cells, were selected from the literature. While these aptamers do not cover the entire repertoire of possible targets, they do represent a substantial portion of the available modified RNA aptamers expected to target cancer cells. Preliminary experiments were run to determine if there was a significant change in the aptamer distribution before and after exposure to cells. The panel of aptamers was mixed in an equimolar proportion and applied to A431 cells at 4 different concentrations,

in duplicate: 0.01pmol, 0.1pmols, 1pmol, and 2pmol per $1x10^6$ cells. Nonbinding aptamers were rinsed off and the cell lysate was collected. Each sample was reverse transcribed then amplified to concentrate the aptamers remaining in the cell lysate. The naive panel, which had not been exposed to cells, was also reverse transcribed and amplified. Each sample was then gel purified to remove unincorporated primers and submitted for next generation sequencing. A principal component (PCA) was performed to determine if the variance in the proportion of each aptamer from the cell lysate was distinguishable from the naïve panel. The results indicated that a distinct difference was observed between the panel and the samples derived from the cell lysate (Figure 3.2). In this plot, is the first component, i.e. the axis that explains the most variance, is the axis that represents this difference.

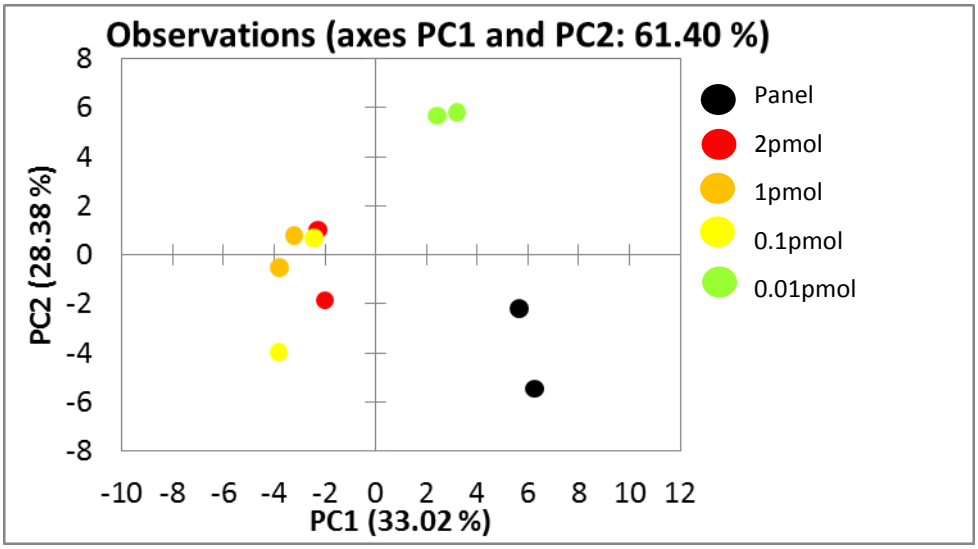



Figure 3.2. PCA comparing various staring panel concentrations to the naïve panel. The F1 axis captures the majority of the variance (33.02%) and accounts for the separation of the panel from the naïve panel. The F2 axis accounts for the separation of the 0.01pmol sample from the others.

In this case, 33.02% of the variance in the model is captured by the first component. It was also observed that starting concentration of the aptamer panel contributes significantly to the variance observed in the aptamer proportions from sample to sample. This is indicated in Figure 3.3, a plot of the first components vs. the third component. Here, 6.34% of the total variance in the data is captured by the third component. While not a large value, it accounts for 10% of the total variance captured by the first three components. In this plot all the samples, save one, are ordered by their concentration.

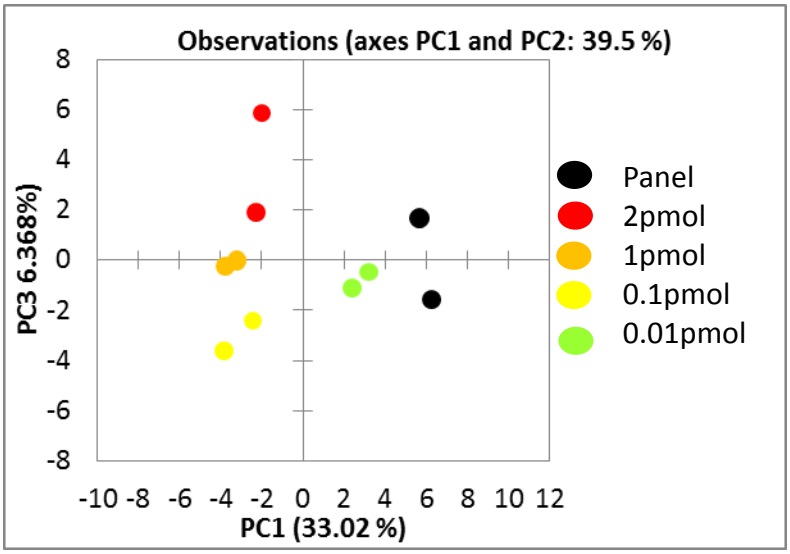



Figure 3.3. PCA plot comparing panel to naïve panel, showing F1 and F3 axis. The F3 axis is dominated by increasing concentration from 1pmol to 2pmols.

The exception to this the 0.01pmol sample. This sample seems to behave differently than the other samples derived from cell lysate. This sample is consistently oriented closer to the panel on the first component, than to the cell lysate samples. This may indicate that the concentration is below what could be considered the limit of detecting and what is

being isolated is residual unbound panel, which would, as observed, be more similar to the naive panel than to the cell lysate panel.

### 3.3. Positive control selection

An examination of the fold change of each of the aptamers revealed that the positive control was not behaving as predicted. Figure 3.4 shows the fold change of each

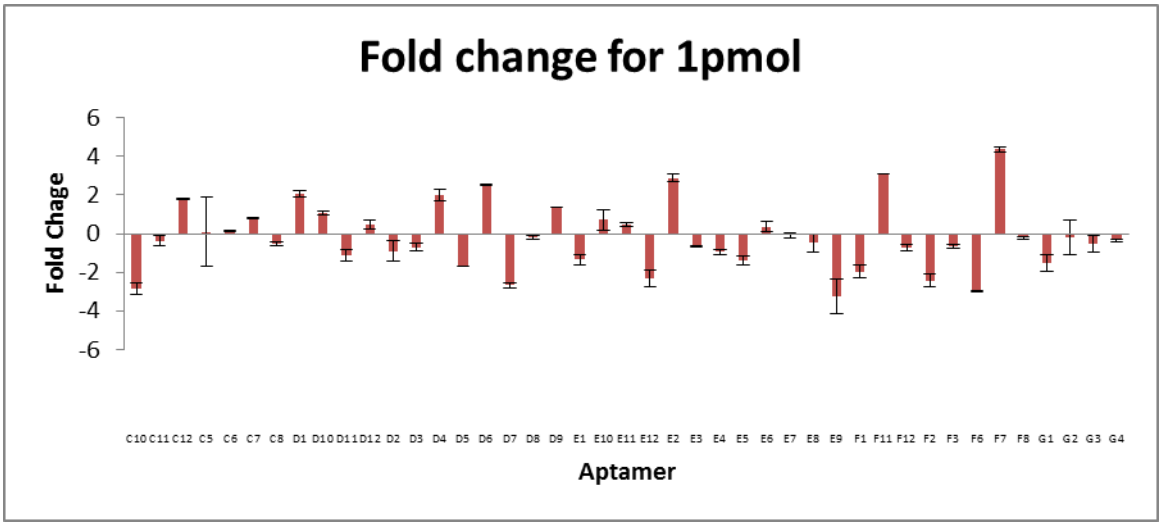



Figure 3.4. Fold change of each aptamer as compared to the naïve panel for the A431 cell line. Fold change is calculated by the $\log_2$ ratio of the abundance of an aptamer from an experimental sample over the abundance of an aptamer from the naïve panel. Error bars represent standard error.

aptamer as compared to the naïve panel. Aptamers with a positive fold change indicate that the aptamer is more abundant in the panel isolated from the cell lysate than in the naïve panel. Aptamers which show a negative fold change are aptamers which are less abundant in the panel isolated from the cell lysate compared to the naïve panel. These

fold changes are correlated with the relative affinity of each aptamer to the target, A431 cells. Aptamer D10 is an aptamer known to target epidermal growth factor receptor (EGFR). The rationale for selecting A431 cells as an initial analysis tool was because they are known to over express EGFR (Ulrich et al., 1984), thus aptamer D10 should show substantial enrichment.

Figure 3.5 shows that the unusual behavior seen for aptamer D10 in not a sequencing error. In a real time PCR experiment, done to verify sequencing results, D10 still showed very little enrichment in the samples. In fact, at the lower concentrations of 0.1 and 0.01pmol D10 actually showed depletion.

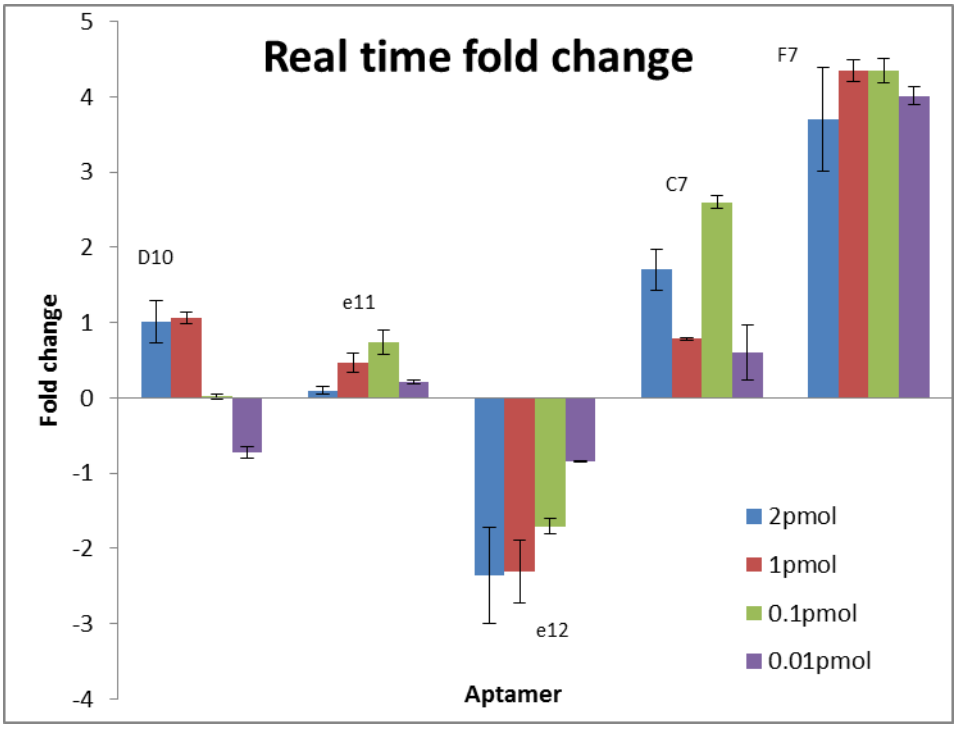



Figure 3.5. Fold change of real time (ΔCt) for representative aptamers, D10, E11, E12, C7, F7 across all concentrations. The fold change is calculated as the $\log_2$ ratio of the Ct a single aptamer from the experimental samples over the Ct of the same aptamer from the naïve panel. Error bars represent standard error.

We predicted that the anomalous behavior seen in aptamer D10 may be due to the modifications made to the sequence of the aptamer to make them compatible with NGS. Forward and reverse primer regions were concatenated to the aptamers in order to facilitate NGS. It is conceivable that these regions perturbed the tertiary structure of the aptamer, thus altering its affinity for the A431 cell line. In order to investigate this, both FACs and real time experiments were performed to estimate the affinity of various aptamers known to bind A431 cells. The aptamers selected to use were, D10 (called e07 in the literature), a minimized version of e07, c1 a ubiquitous internalizing binder developed by Levy, otter an internalizing aptamer developed by the Ellington lab, and C36 a negative control aptamer (Magalhães et al., 2012).

For the FACs study each aptamer was folded with one or two oligonucleotide complementary to the concatenated primer regions. The complementary oligonucleotide at the 5' end was conjugated through biotin to phycoerythrin to allow visualization of the aptamer bound to a cell. Figure 3.6 shows the results of this analysis. Minimized e07 and otter showed little signal of over background; their mean fluorescence is only modestly higher than that of C36, the negative control. C1 and e07 clearly show affinity for the cell line and in each case having both the 5' and 3' primer regions blocked by complementary oligonucleotides shows superior affinity. Because this FACs analysis relied on a reporter conjugated to a complementary oligonucleotide to show signal, it was not possible to test having only the 3' region blocked. In order to test this, real time

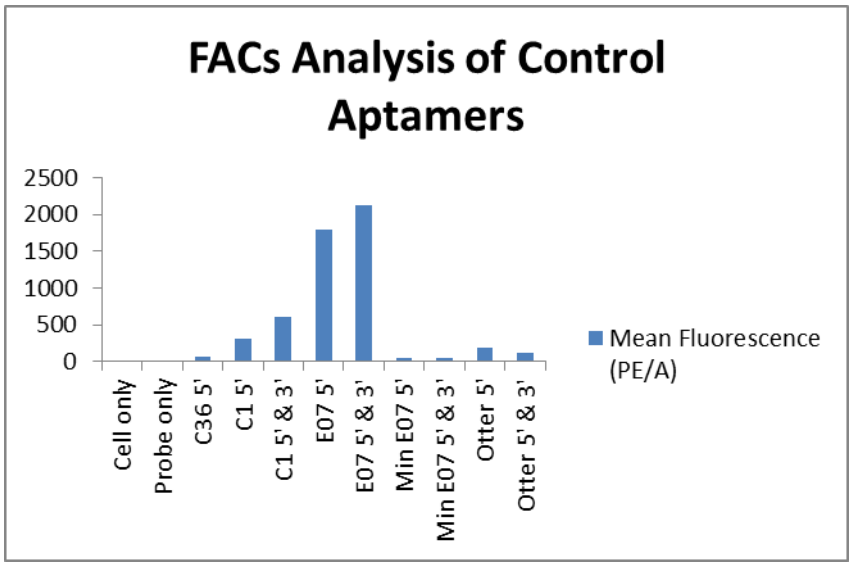



Figure 3.6. Results of FACS analysis of 4 possible positive control aptamers. Cell only, probe only and C36 are negative controls and are not expected to show fluorescence. Figure courtesy of Michelle Byrom.

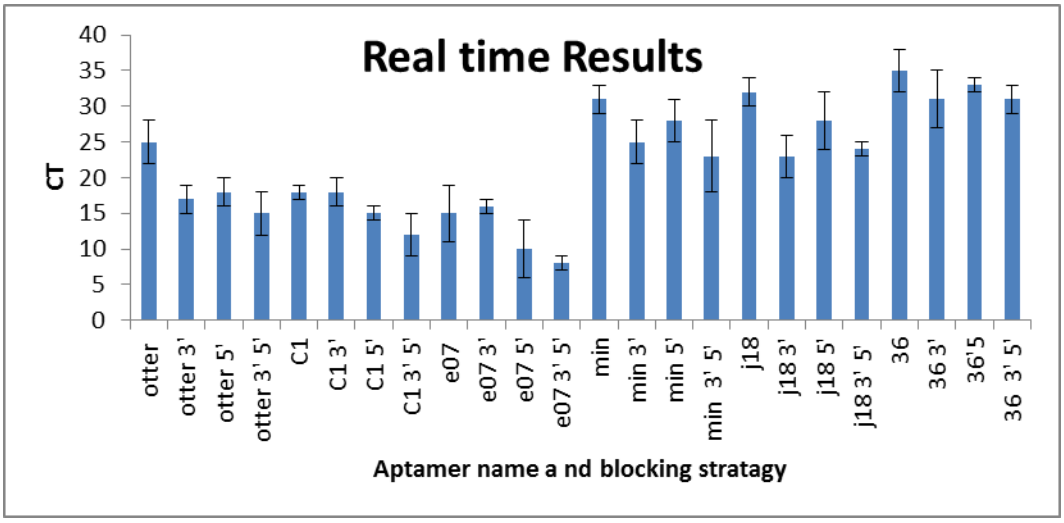



Figure 3.7. Real time PCR analysis. Each aptamer, otter, C1, e07, min e07, j18, and 36, was tested with either the 5' constant region, the 3' constant region, both or neither region blocked by complementary oligonucleotides. Aptamers with a strong affinity for the target will be more abundant in each of the samples and have a lower Ct. Error bars represent standard deviation.

PCR was used to estimate the abundance of various aptamers after cell binding. As with the FACS experiment, each aptamer was folded with an oligonucleotide complementary to either the 5' region, 3' region, both or neither. J18 an EGFR RNA aptamer developed by the Ellington lab (Li et al. 2010) was tested in addition to the aptamers tested in the FACS experiment.

Figure 3.7 shows the results of this study. As previously observed, minimized e07 and otter showed the lowest affinity for the A431 cell line; as evidenced by the high Ct for each of the samples. J18 also displayed poor performance. In nearly all cases, blocking both the 3' and 5' regions shows superior binding affinity. Furthermore, blocking only the 5' region showed better affinity than blocking only the 3' region, which implied that the 5' constant region has more effect on the structure than the 3' region though both play a roll. Table 3.2 shows the difference between the Ct for the unblocked aptamer and the Ct for each of the blocking strategies plus the difference between the blocked aptamer and a non-binding aptamer, the largest values correlate with the most improvement over the unblocked aptamer. The largest values are for 5' blocked e07, 3'5' blocked e07, 3'5' blocked otter and 3'5' C1. Table 3.3 shows a statistical of the significance of the difference between the blocked and unblocked aptamers. All of the aptamers that show the largest values in Table 3.2 show a significant difference. Considering the results from the FACs study and the real time study it was decided to consider e07 the positive control.

| | 3' | 5' | 3',5' |
|---|---|---|---|
| Otter | 22 | 22 | 26 |
| C1 | 13 | 21 | 25 |
| e07 | 14 | 28 | 30 |
| min | 12 | 8 | 16 |
| J18 | 17 | 9 | 15 |

Table 3.2 Sum of the difference between blocked / unblocked aptamers and between blocked aptamers / non-binder.

| | 3' | 5' | 3',5' |
|---|---|---|---|
| otter | 0.02 | 0.03 | 0.02 |
| C1 | 1 | 0.02 | 0.03 |
| e07 | 0.7 | 0.2 | 0.04 |
| min | 0.04 | 0.2 | 0.06 |
| J18 | 0.01 | 0.2 | 0.003 |

Table 3.3 Significance of the distance between blocked and unblocked aptamers.

### 3.4 Analysis of cell lines A431 and MDA-MB-435

We wanted to determine if the variation observed in the cell lysate samples was cell specific. A431 cells and MDA-MB-435 cells were used as the trial cell lines. These were selected because A431 over expresses epidermal growth factor receptor (EGFR) and MDA-MB-435 is considered to be negative for EGFR. The panel being used

contained an aptamer specifically targeted to EGFR and could be used to lend biological rational to the orientation of the cell samples on the PCA plot.

A431 cells and MDA-MB-435 were incubated with 4 different concentrations of the panel, in duplicate: 0.01pmol, 0.1pmol, 1pmol, and 2pmol per $1x10^6$ cells and collected the lysate as described above. After NGS it was observed that there did seem to be a trend separating the A431 cells from the MDA-MB-435 cells (Figure 3.8). In all cases, save the 0.01pmol sample, the MDA-MB-435 cell lines are higher than the A431 cells on the F2 axis.

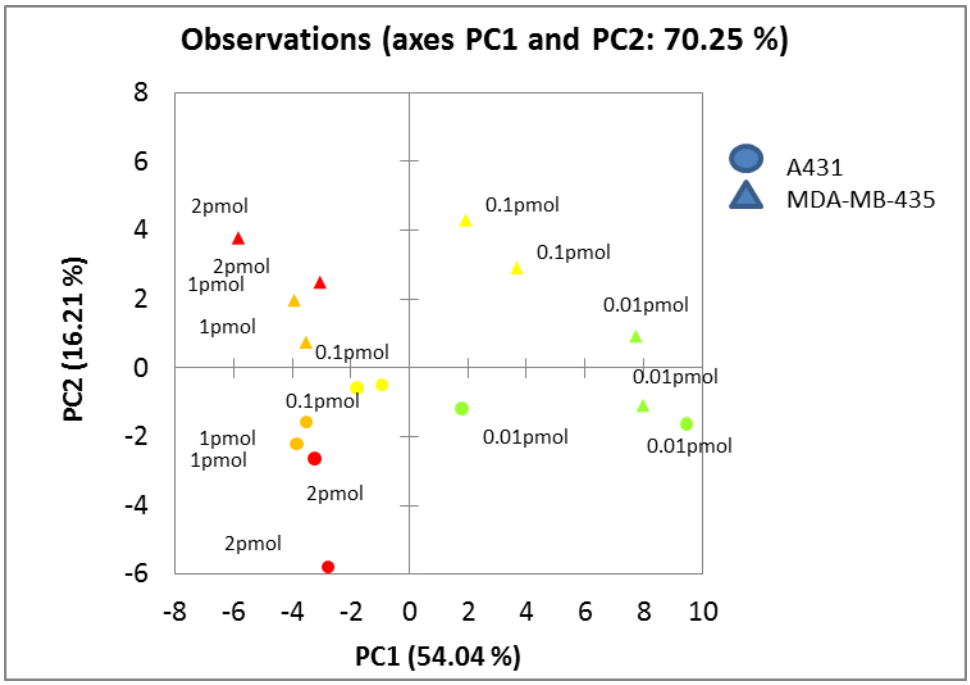



Figure 3.8. A PCA plot comparing two cell lines A431 and MDB-MB-435. The F1 axis is dominated by the starting concentration with low concentrations being on the left and high concentrations being on the right, . The F2 is dominated by the between cell line differences. Most of theMDA-MB-435 cells are above the axis while most A431 cells are below the axis.

Unfortunately the F1 axis was still dominated by the starting concentration of each of the samples. From the loading plot (Figure 3.9) it appears that most of the aptamers are important to the F1 axis (indicated by being more parallel with the axis) and thus their behavior may be more strongly influenced by concentration rather than cell to cell differences.

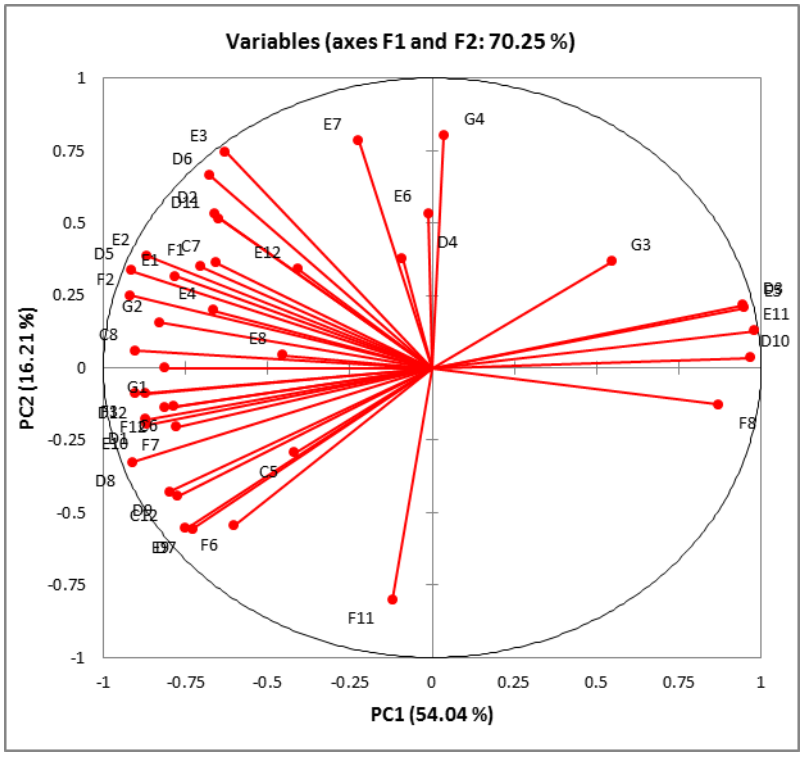



Figure 3.9. Correlation plot relating the significance of each variable to the position of the cells on the plot. Note that D10 (the EGFR specific aptamer) seems to behave in a concentration dependent manner and contributes more to the separation based on concentration rather than cell type. Aptamers G4 and F11 seem to play the most important role in discriminating the two cell lines.

In an attempt to determine which aptamers were most resistant to concentration effects, a PCA was rerun for each of the different sample concentrations. Figure 3.10 shows the results of this analysis. Because concentration is no longer a factor, the F1 axis captured information regarding cell to cell differences. As previously observed, the

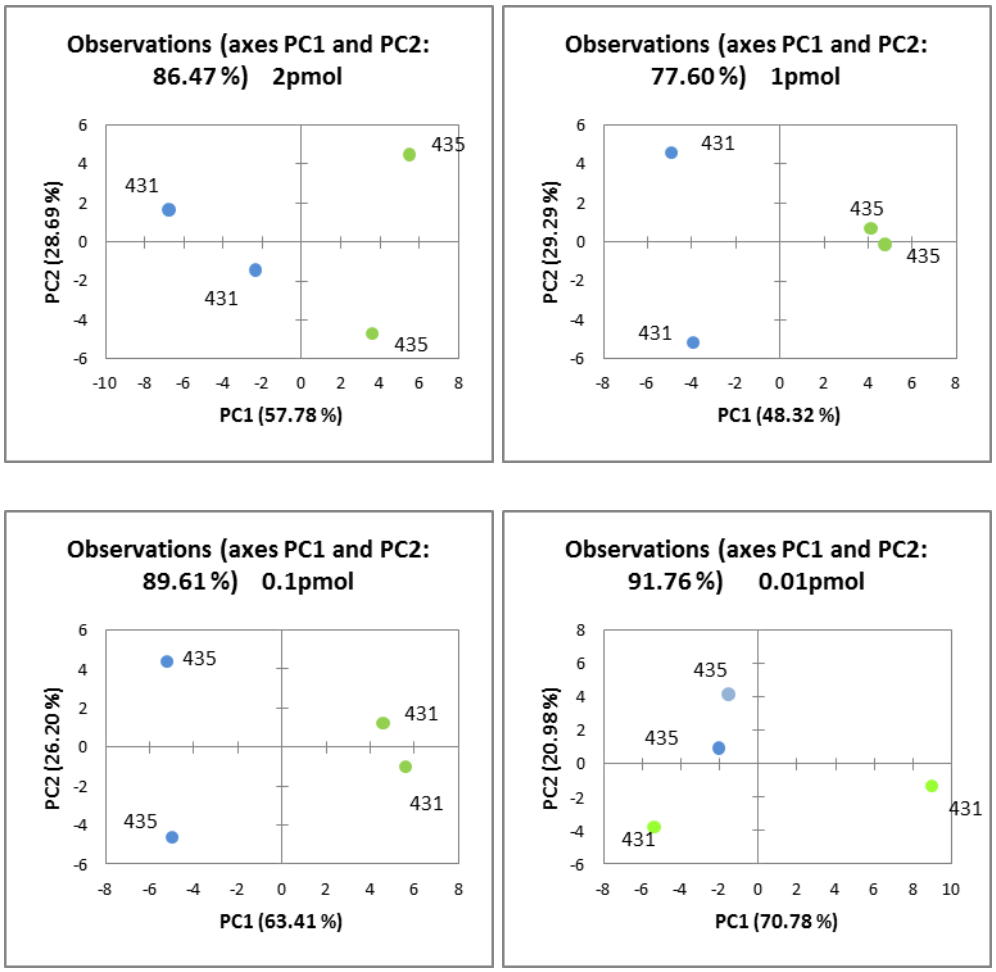



Figure 3.10. PCA analyses of each sample concentration independent from all others. In each case, save 0.01pmols, the F1 axis is the axis of separation separating the A431 cell line of the right from the MDA-MB-435 cell line on the left. For the 0.01pmol sample the axis of discrimination is the F2 axis separating A431 on the bottom from MDA-MB-435 on the top.

0.01pmol sample behavior deviates from the behavior of the other samples. In this case the F1 axis is dominated by difference between cells of the same type and cell to cell differences are captured on the F2 axis. This indicates that while there is some difference in the behavior of the aptamer panel due to cell to cell differences, the behavior is obscured by background differences in individual cultures.

The 0.1pmol, 1pmol and 2pmol data was used to identify which aptamers may be most important to cell line discrimination. For each PCA analysis a correlation table was generated which indicates the level of correlation each aptamers has with each axis. For each cell sample, aptamers which had a correlation value of at least ±0.75 were identified (Table 3.4). These aptamers were used in a new PCA analysis including all sample concentrations. Figure 3.11 shows the results of this analysis. The F1 axis now relates to the cell to cell variation. Nevertheless different starting concentrations did affect the behavior of each of the aptamers as indicated by the amount of variance captured by the F2 axis, 23.77%

| 2pmol | 1pmol | 0.1pmol | 2pmol | 1pmol | 0.1pmol | Consensus |
|---|---|---|---|---|---|---|
| D2 | C12 | C12 | E9 | F6 | E2 | D2 |
| D3 | D1 | C5 | E11 | F7 | E3 | D3 |
| D5 | D2 | C6 | E12 | F8 | E4 | D6 |
| D6 | D3 | C8 | F1 | F11 | E5 | D7 |
| D7 | D4 | D2 | F2 | | E6 | D8 |
| D8 | D5 | D3 | F6 | | E8 | E2 |
| D9 | D6 | D4 | F8 | | E9 | E3 |
| D11 | D7 | D5 | F11 | | E10 | E9 |
| E1 | D8 | D6 | G4 | | E11 | F6 |
| E2 | D10 | D7 | | | F3 | F11 |
| E3 | D10 | D8 | | | F11 | |
| E5 | E1 | D9 | | | G1 | |
| E6 | E2 | D10 | | | G2 | |
| E7 | E3 | D11 | | | G4 | |
| E8 | E9 | D12 | | | | |
| Continued in next columns | | | | | | |

Table 3.4. List of aptamers with a correlation value of at least ±0.75 across the F1 axis for sample concentrations: 0.1pmols, 1pmols, 2pmols. The consensus Colum contains the aptamers which have a correlation value of at least ±0.75 for all concentrations.

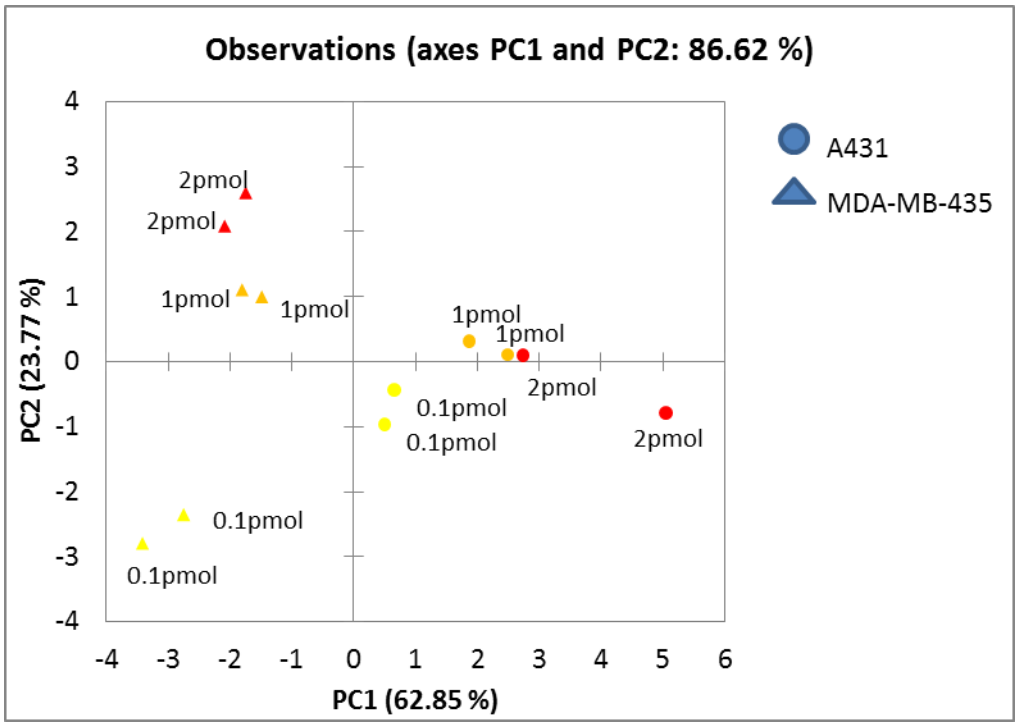



Figure 3.11. PCA of the two cell lien data, using only the consensus values recorded in Table 3.1, excluding data from the 0.01pmol sample. In this figure the F1 axis contains data relevant to cell type, all A431 samples are found on the right and all MDA-MD-435 samples are found on the left. Concentration remains as a significant source of variation.

### 3.5 Analysis of four cell lines, A431, MDA-MG-435, Hek and U87MGvIII

In order to overcome noise derived from starting panel concentration and fully validate this method, a single concentration was selected for further analysis, 1pmol per $1x10^6$ cells. Based on the results seen in Figure 3.10, it appears that the aptamers in the 1pmol and 2pmol concentrations are more resistant to variations in concentration than the aptamers in the 0.1pmol concentration, on account of the proximity of each of the points

to each other. Two pmol was not selected in order to conserve reagents. Four different cell lines were selected for this experiment. The A431 and MDA-MB-435 pair was used. A glioma line, U87MGvIII, was selected because its tissue type was very disparate from the other lines used and it expresses a mutant strain of EGFR, to which aptamer D10 also binds. The HEK line was selected for being a disparate tissue type and

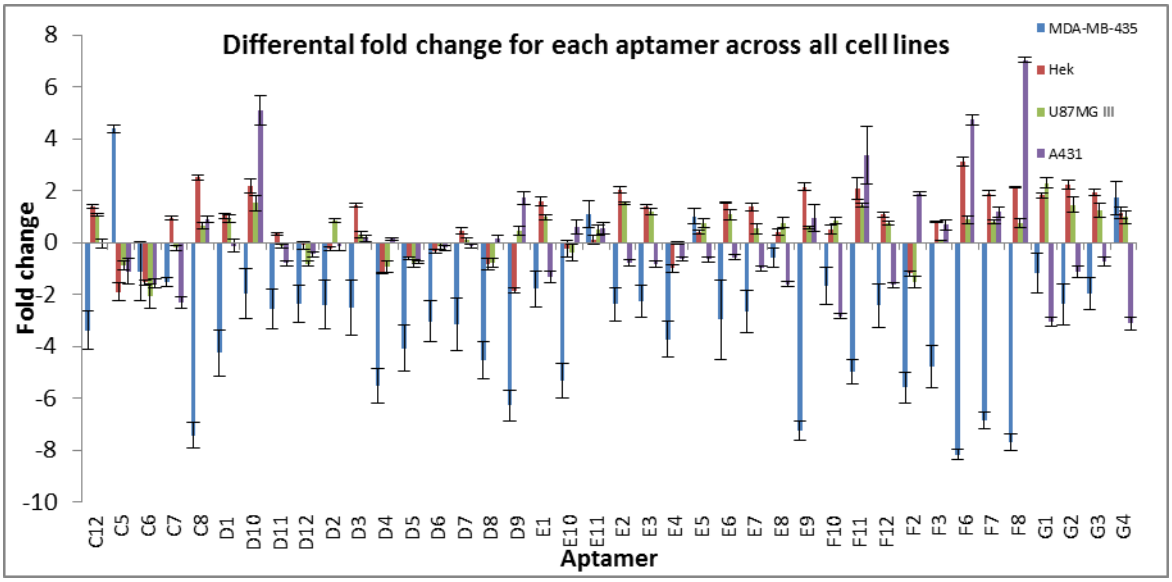



Figure 3.12. This plot shows that each cell line has a unique pattern also shows that negative values ( as the case in F11 and G4) can also be significant. To include or not to include that is the question.

to represent a “normal” cell line. The experiment was performed as previously described in quadruplicate for each of the cell lines. Figure 3.12 shows the fold change for each of the aptamers as compared to the naïve panel. It is clear from this figure that the aptamers behave differently depending on which cell line they are exposed to. For example very few aptamers seem to bind to the MDA-MB-435 cell line. The HEK and U87MGvIII line

on the other hand seen to have very similar aptamer binding profiles. Despite this there were subtle differences in the aptamer binding patterns for these two cell lines.

After performing a PCA, four distinct groupings were observed. (Figure 3.13) Each grouping is distinctly separated from the other lines. The F1 axis is primarily

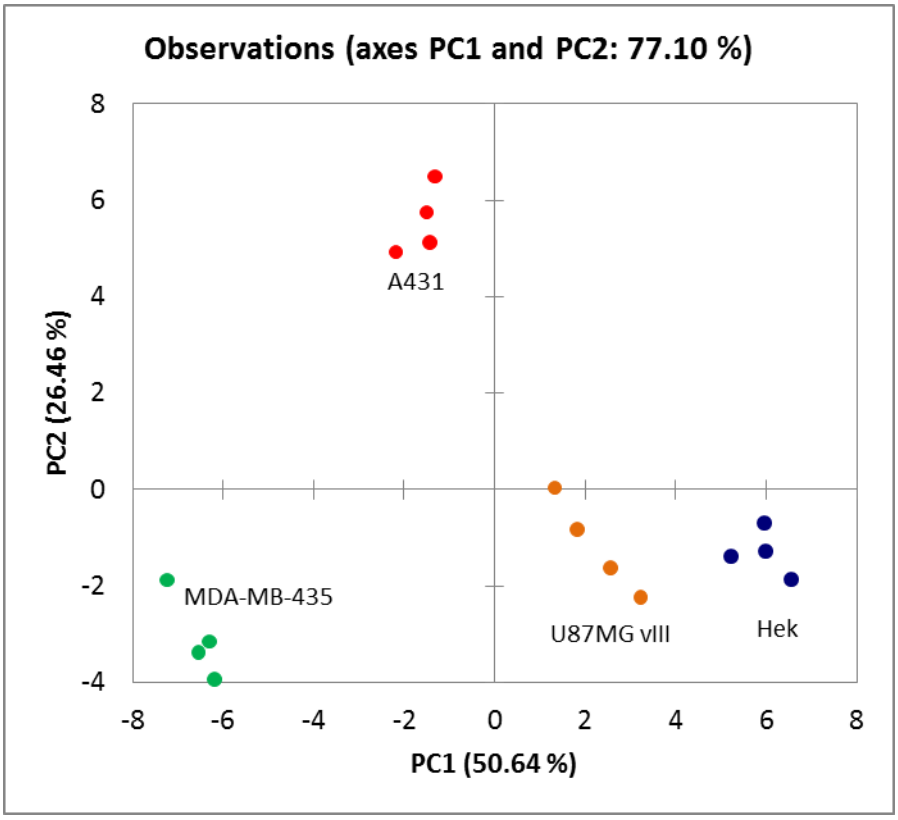



Figure 3.13 PCA of four cell line experiment. Four distinct groupings are observed, HEK, MDA-MB-435, A431 and U87MGvIII. The F1 axis represents the variation which separates MDA-MB-435, U87MGvIII and HEK cell lines. The F2 axis represents the variation that separates the A431 cell line from the other lines. Additional separation of the U87MGvIII is found on the third axis (not shown).

responsible for the separating between the MDA-MB-435 cell line and the Hek cell line.

A431 is primarily discriminated across the F2 axis. U87MGvIII is primarily discriminated across the F3 axis. When the correlations between the aptamers and the axis are examined (Table 3.3) several interesting associations are observed. First, D10 was found to play a role in the location of the A431 cell line, as expected since A431 over expresses EGFR and D10 is known to bind EGFR.

| F1 | F2 | F3 |
|---|---|---|
| **C12** | D4 | **D2** |
| **C5** | D8 | |
| C7 | D9 | |
| C8 | D10 | |
| D1 | E10 | |
| D11 | **E5** | |
| D12 | **F2** | |
| D3 | F6 | |
| D5 | F8 | |
| D7 | G4 | |
| E1 | | |
| E12 | | |
| E2 | | |
| E3 | | |
| E6 | | |
| E7 | | |
| F1 | | |
| F12 | | |
| F7 | | |
| G2 | | |
| G3 | | |

Table 3.5. List of aptamers with correlation values ±0.75 for the first three axes. The values with the highest and lowest correlation values for each of the columns are in bold.

While Hek and U87MGvIII also express EGFR (or the mutant version in the case of U87MGvIII), they do so at levels less than that of A431. C12 had the highest correlation value for the F1 axis at 0.98; this aptamer was selected against prostate specific membrane antigen (PSMA). This implies that the cell lines Hek and to a lesser extent, U87MDvIII may carry molecules similar to PSMA. C5, an aptamer selected against plasminogen activator inhibitor-1 (PAI-1), had the strongest negative correlation of -0.81. Along the second axis, F2 had the highest correlation of 0.97 and is an aptamer selected for cell surface CD4. Conversely, E5 had the strongest negative correlation of -0.86 and was selected against G coupled neurotensin receptor (NTS1). Finally, the only aptamer which showed high correlation with the F3 axis was D2 with a correlation value of 0.80. This aptamer was selected against H526 cells, a small cell lung carcinoma line.

### 3.6 Validation of model

To fully validate this model a discriminate analysis (DA) was performed. This analysis will seek the axis of rotation that best maximizes the distance between groups while minimizing the distance between members of the same group. The form of PCA used in this manuscript separates the groups based entirely on the variance in the dataset, while it is important to know that variance plays a large role in the model, it does not mean that variance alone accounts for group separations. A DA has the ability to groups samples based on characteristics other than variance, though it often unclean what those characteristics are. Figure 3.14 shows the results for the DA. The DA was able to find

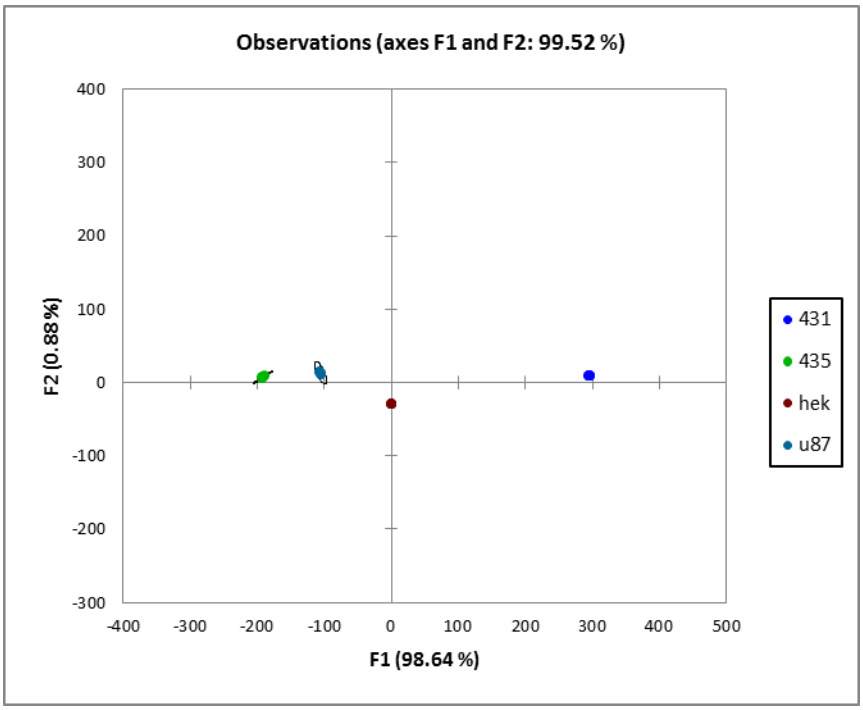



Figure 3.14 Discriminant analysis of 4 cell lines. Each group is clearly separated from each other group across the F1 axis. 98.64% of the data in the model exists across the F1 axis; this is the between group scatter. Data relating to within group scatter is found on the F2 axis and only accounts for 0.88% of the data in this model.

an axis of rotation that grouped members of the same class extremely close to one another while maintaining the separation between the groups. The Fischer distances in Table 3.6 underpin the notion that each group is a distinct entity. The results of a Fischer test to assess the uniqueness of the groups shoed a p-value of less than 0.0001 for each pair of cell types (Table 3.7).

A leave-one-out cross validation analysis excluded each point from the data set iteratively and re-analyses the remaining data. The method then tries to predict the class

| | 431 | 435 | Hek | u87 |
|---|---|---|---|---|
| 431 | 0 | 17599.476 | 6557.710 | 11984.413 |
| 435 | 17599.476 | 0 | 2872.844 | 641.080 |
| hek | 6557.710 | 2872.844 | 0 | 999.721 |
| u87 | 11984.413 | 641.080 | 999.721 | 0 |

Table 3.6. Pairwise Fischer distances.

| | 431 | 435 | Hek | u87 |
|---|---|---|---|---|
| 431 | 1 | < 0.0001 | < 0.0001 | < 0.0001 |
| 435 | < 0.0001 | 1 | < 0.0001 | < 0.0001 |
| hek | < 0.0001 | < 0.0001 | 1 | < 0.0001 |
| u87 | < 0.0001 | < 0.0001 | < 0.0001 | 1 |

Table 3.7. Pairwise Fischer distance tests.

of the excluded point. For this data set each point was correctly classified (Table 3.7). Upon examination of the classification functions, it was found that only 9 aptamers were required for classification by this model (Table 3.9). Aptamers C5, C12 F2 and D2 were 4 of the 5 aptamers that showed the strongest correlation in the PCA model. Aptamer E5, the aptamer with the strongest negative correlation to the F2 axis in the PCA model, was not required for classification by the DA. Curiously of the reimaging aptamers, all were considered important for classification in PCA, save aptamers E11 and F10.

| from \ to | 431 | 435 | Hek | u87 | Total | % correct |
|---|---|---|---|---|---|---|
| 431 | 4 | 0 | 0 | 0 | 4 | 100.00% |
| 435 | 0 | 4 | 0 | 0 | 4 | 100.00% |
| hek | 0 | 0 | 4 | 0 | 4 | 100.00% |
| u87 | 0 | 0 | 0 | 4 | 4 | 100.00% |
| Total | 4 | 4 | 4 | 4 | 16 | 100.00% |

Table 3.8 Results or cross-validation.

| Classification aptamers |
|---|
| C5 |
| C12 |
| D2 |
| D3 |
| E10 |
| E11 |
| F2 |
| F6 |
| F10 |

Table 3.9 List of aptamers used for classification by the DA.

### 3.7 Exploration of aptamer behavior as a panel and alone

Over the course of these experiments certain trends were observed in aptamer behavior as a function of concentration and affinity for a target. For example, D10, the EGFR aptamer, seemed to become more enriched in the experimental samples when the

total aptamer starting concentration was the lowest. At the highest concentration, 2pmol, D10 barely had a 2 fold enrichment over the naïve panel (Figure 3.15).

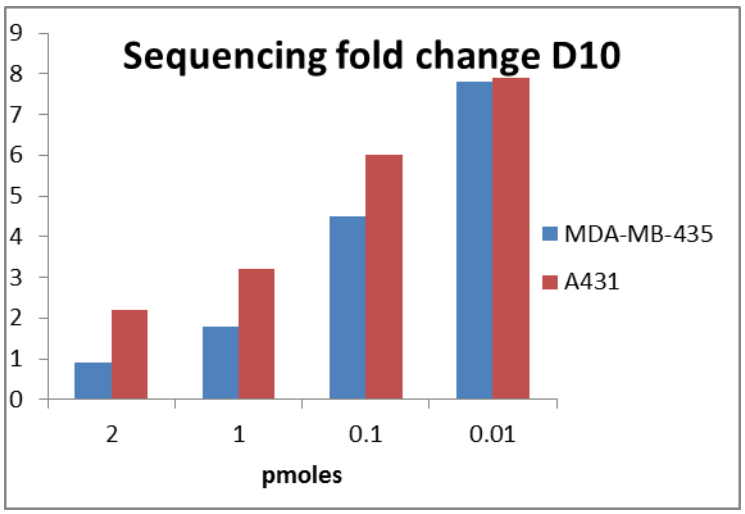



Figure 3.15. Fold change of aptamer D10 calculated from sequencing data for the A431 and MDA-MB-435 cell lines.

C7 on the other hand is an aptamer selected to target a fragment of mouse VCAM-1. In general, it showed depletion or minimal enrichment at all concentrations for the A431 and MDA-MB-435 cell lines. D10, which could be considered a "good" aptamer, showed its best affinity for both cell lines at the 2pmol concentration (Figure 3.15).

In order to investigate this further it was decided to perform a FACs analysis to explore the affinity of each aptamer for each cell line alone or as part of a panel. One pmol of labeled aptamers D10 or C7 were incubated with each of the experimental cell types, A431, Hek, MDA-MB-435 or U87MGvIII. Each labeled aptamer was also mixed with 1pmol of each other aptamer in the panel and incubated with the various cells. A

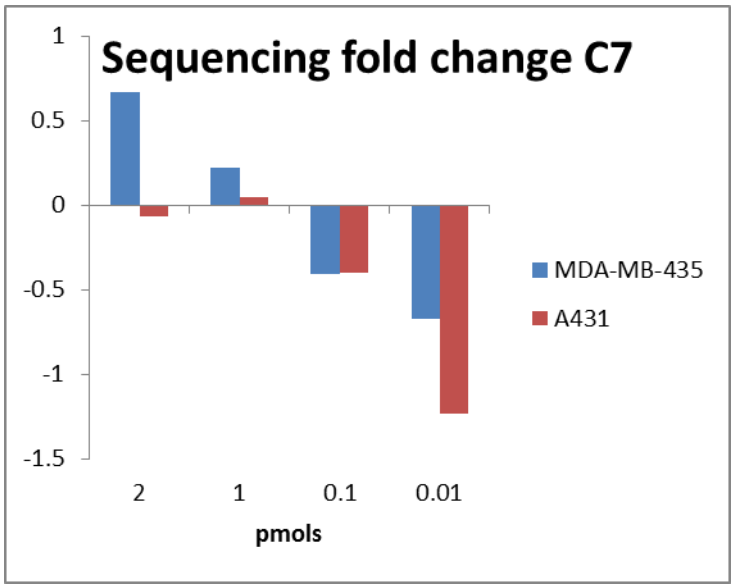



Figure 3.16 Fold change of aptamer C7 calculated from sequencing data for the A431 and MDA-MB-435 cell lines.

10pmol positive control was included for each aptamer, as 10pmol is the standard concentration used for FACs with respect the aptamer D10 and the A431 cell line. The negative controls included cell only, label only, and H12, a randomly described oligonucleotide. Figure 3.17 shows the results of this assay for aptamer D10. In each case the aptamer as part of the panel shows greater affinity for the cell line than the aptamer by itself. As expected in three of the cases the 10 pmol aptamer concentration showed high fluorescence than the 1 pmol concentration. The notable exception to this is the MDA-MB-435 cell line. This line is considered an EGFR negative line (recall that D10 binds EGFR) and thus D10 could be considered a "bad" aptamer for this line. Though the response seen for this line is only modestly above back ground it is interesting to note that the 1 pmol aptamer concentration as a panel performs better than the 10 pmol concentration.

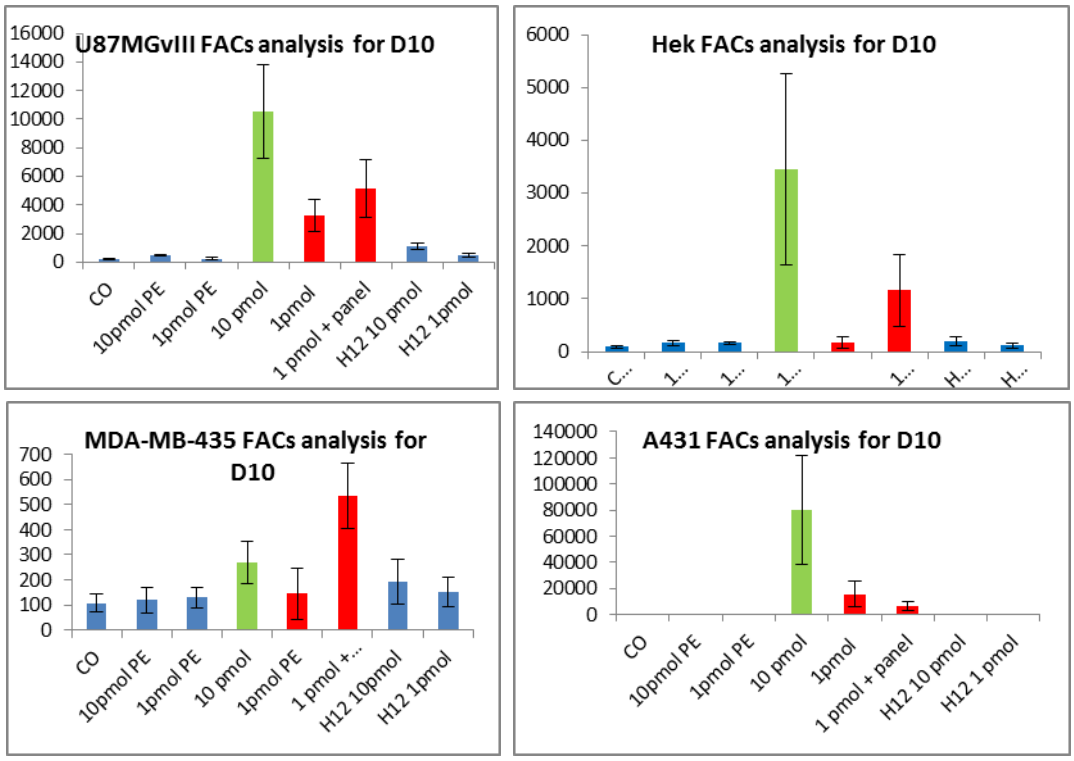



Figure 3.17 FACs analysis of aptamer D10 under various conditions for each of the cell types. In all cases, save for the A431 cell line, aptamer D10 shows greater affinity for the cell line as part of a panel that alone.

C7 is an aptamer which showed very little affinity for any of the cell lines based on sequencing fold change (Figure 3.12). Figure 3.18 shows the results of a real time analysis for this aptamer. As expected the response in only marginally over baseline yet interesting trend are still observed. As with D10 the aptamer as part of a panel shows greater affinity for the cell line than the aptamer alone, albeit much more modestly than for D10. It is also observed that the 1 pmol aptamer concentration as a panel performs better than the 10 pmol concentration for the MDA-MB-435 cell line; the cell line that showed the greatest depletion in the sequencing results.

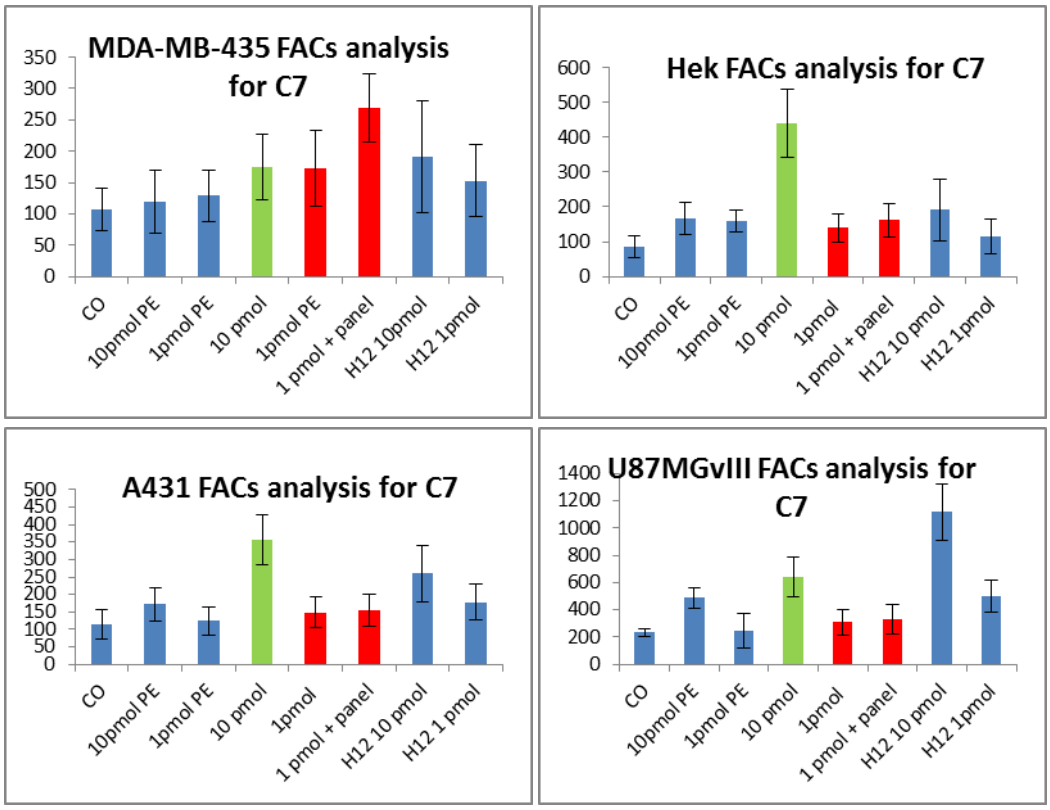



Figure 3.18. FACs analysis of aptamer C7 under various conditions for each of the cell types. In all cases aptamer C7 shows a very slight increase in affinity for the cell lines as part of a panel that alone. It should be noted however that much of this data is not above background and should be viewed with that 0in mind.

# 4 Discussion

In this work, the utility of aptamers as non-specific receptors for biomolecular discrimination was explored. We have shown that a modest panel of aptamers can accurately discriminate at least four different cell lines. Unlike other methodologies, this approach does not rely on a fluorogenic signal and thus overcomes the limitations of fluorescence experiments. Using NGS technology in a novel manner has allowed us to utilize relative abundances of aptamer concentration as a receptor signal to create a

unique fingerprint of aptamer binding without the need to perform separate experiments to assess the affinity of each aptamer for each target. By exploiting properties of nucleic acids which allow simple sequencing, we have functionally created a detecting platform bounded only by the number of sequences a researcher choses to use. This is in stark contrast to other platforms which rely on multiple visual signals to interrogate results or special location on a solid support to increase the dimension of interrogation.

Aptamers, by virtue of being synthetic nucleic acids, are the ideal choice to exploit the power of NGS to generate vast amounts of data. They are simple to create de novo in very large quantities and can be selected or designed to display many different chemical behaviors. In this study, the aptamers used were selected by other researchers to target cells or proteins which would be expected to be present on the surface of cells. Presumably, one could achieve results similar to those that we have presented here by creating one aptamer specific to each cell line. However, the sheer number of targets that one would need to design to discriminate many different cell types, makes the idea of one aptamer, one target untenable. Rather, the significance of what we have shown here is that aptamers need not be specific for a single target in order it achieve discrimination.

In this study, a number of aptamers behaved in a predictable manner based on the targets they were selected for. D10 for example was selected for EGFR and it played an important role in discriminating A431, which highly expresses EGFR, from the other cell types. C5 is an aptamer selected to target PAI-1, a serein protease inhibitor associated with tumor malignancy (Madsen 2010). PAI-1, has a strong correlation with breast cancer outcome and when activated, its deposition in the extracellular matrix of cells

facilitates cell motility (Rougier et al., 1998, Look et al., 2002). The correlation of this aptamer with the MDA-MB-435 cell line is not surprising as the line is in fact a human breast carcinoma.

The behavior of aptamer C12 is an interesting case where it the aptamer's behavior is strongly correlated with the F1 axis yet there are two cell lines that seem to interact with the aptamer in a significant manner. This aptamer was selected against PSMA and in the context in which it was selected, seemed to bind in a specific manner (Chu, 2006). Although PSMA is overexpressed specifically in prostate cancers, (Horoszewicz et al., 1987) it is nevertheless found to be expressed in other tumor types such as kidney, glioma and breast (Chang et al., 1999). C12 shows enrichment for the Hek and U87MGvIII cell lines but its abundance remains steady for the A431 cell line and showed depletion for the MDA-MG-435 cell line (Figure 3.11). It may be that the Hek and U87MGvIII cells lines express PSMA at higher levels than A431 and MDA-MB-435 expresses very little PSMA. This is an excellent example of how differing levels of target expression drive cross-reactive binding of an aptamer.

While U87MGvIII can be discriminated across the F1 axis the F3 axis also explains a fair amount of variance between the U87vIII cell line and the other lines. The only aptamer to show strong correlation to the F3 axis was aptamer D2. This is an unexpected result as aptamer D2 was selected against the H526, a small cell lung carcinoma line (Lee, 2007). While cross reactivity of aptamers was the goal of this project it is surprising that two such disparate tissue types should share something in common.

Another surprising result was the affinity of aptamers F11 and F12. These aptamers were selected against Plasmodium falciparum (malaria) erythrocyte membrane protein 1 (PfEMP1) (Barfod et al., 2009). It is striking that an aptamer targeting a protist would be able to also target human cell lines. Aptamer F11 in particular, showed enrichment for all cell lines except MDA-MB-435. In fact was the 4$^{th}$ most enriched aptamer behind F6 (targets CD4 antigen (Kraus et al., 1998)), F8 (targets αvβ3 integrin (Mi et al., 2005) and D10. This lends credence to the idea that aptamers can be cross-reactive with unexpected targets. Unlike antibodies, which rely on very specific chemical interactions between the antibody residues and its target, aptamers seem to be capable of binding many disparate targets, possibly due to their simple chemistries.

While the repertoires of chemical interactions which can be carried out by aptamers are limited, their simplicity could actually be a significant advantage over other receptors. It can be reasonably expected that the exact composition of a complex target would result in slight perturbations of the binding efficiency of each aptamer, thus aiding in generating a unique binding profile for each target. The simplicity of nucleic acids has further benefits in that there is a wealth of non-natural modifications which could be used to "tune" an aptamer panel to particular molecular targets, for example boronic acid modification for saccharide affinity (Edwards et al., 2007). The great versatility displayed by aptamers theoretically can be used for any complex target, including non-peptide molecules and even those of non-biologic origin. For example, there are a number of aptamers developed molecules like cocaine, ATP; aminoglycosides, metallic ions, and porphyrins (Stojanovic et al., 2001; Sazani et al., 2004; Walter et al, 1999; Li et

al. 1996; Zhang et al, 2011) Furthermore, there is no limit to the number of unique patterns which could be generated using aptamer distribution as a signal. This means this method could be a single assay capable of discriminating any complex target. This opens up a new class of receptors for use in differential sensing routines. When used in conjunction with NGS methodologies this technique has the potential to rival mass spectroscopy for complex target discrimination.

# Chapter 4: The Use of Principal Component Analysis and Discriminant Analysis in Differential Sensing Routines

*This chapter is derived from "The Use of Principal Component Analysis and Discriminant Analysis in Differential Sensing Routines" by Sara Stewart, Michelle Adams-Ivy, and Eric Anslyn. Chem. Soc. Rev. 43.1(2013) 70-84*

## 1. Introduction

Differential sensing has become an increasingly important concept in the field of supramolecular chemistry, as trends in research shift from using lock-and-key receptors to employing less selective receptors in array sensing (Lavigne, et al., 2001; Anslyn, 2007; Collins and Anslyn, 2007; Miranda, et al., 2010; Musto and Suslick, 2010; Umali and Anslyn, 2010) Modeled after the mammalian olfactory senses, differential sensing employs a collection of low selectivity receptors that signal a specific pattern for each analyte or complex solution. In turn, each analyte or solution is discriminated from others by a unique fingerprint. In practice, the fingerprint, consisting of various fluorescence, absorbance, or electrode data cannot be easily analyzed by individual calibration curves for the purpose of analyte and solution identification and differentiation.

To alleviate such difficulties, chemists have explored the use of statistical analysis techniques such as principal component analysis (PCA) and discriminant analysis (DA). Although these techniques are becoming particularly important for differential sensing purposes, (Hirsch et al. 2003; Buryak and Sevenn, 2005(2); Zhou et al. 2006; Palacios,

2007; Zang and Suslick, 2007; Hughes et al., 2008; Shabbir et al., 2009; Bajaj et al., 2010(2); Takeuchi et al. 2011) these techniques are sometimes used as a "black box". PCA and DA are widely utilized across multiple fields of academia and industry, thus there are numerous reviews and tutorials on these techniques available to study (Klecka, 1980; Fukunaga, 1990; Coomans and Massart, 1992; Lewi, 1992; Jurs er al., 2000; Hardle and Simar, 2003; Iznman, 2008; Theodoridis, 2009). However, these articles are often heavily laden with mathematical symbols and derivations, or with seemingly unrelated examples, that are challenging to translate to differential sensing. For this reason, we see the present need for a qualitative explanation of these techniques to help chemists interpret PCA and DA plots that result from differential sensing studies. Our aim is to present PCA and DA to chemists in a manner that will shed light on the types of receptor arrays that lead to certain plots, and to give a few general criteria for obtaining optimal PCA and DA plots. This information ultimately can be utilized to refine differential sensing systems for better analyte and solution discrimination and differentiation.

## 2. Background

Both PCA and DA are statistical analysis techniques, that produce score plots for the analytes or solutions tested. These score plots consist of a coordinate system utilizing axes in a two, three, or higher dimensionality space, with the goal of revealing the coordinate system in that the test analytes are best discriminated. Both PCA and DA generate these score plots by decomposing the raw data by a matrix technique, in that the

Eigenvectors of the matrix produce axes mentioned above and the eigenvalues give a measure of the level of discrimination that exists in the data. However, the manner in that each of these techniques arrives at their corresponding Eigenvectors and Eigenvalues is slightly different.

To explore how PCA and DA work, a relevant analogy can be made to a more familiar Eigenvalue problem (see Scheme 4.1 for how Eigenvalue problems are written). Most chemists know that the Schrödinger equation plays a fundamental role in quantum mechanics (Equation 4.1). This differential equation is usually simplified and reduced to a problem involving the Eigenvectors and Eigenvalues of a square matrix (Korn and Korn, 1961). The Eigenvectors of this matrix represent the molecular orbitals with that we are all familiar, and the Eigenvalues give the orbital energies that correspond roughly to the ionization potentials of the molecules. This classic equation is just one of many examples of Eigenvalue problems, that play roles in fields as diverse as signal processing and civil engineering. Simply stated, when a matrix is multiplied by one of its Eigenvectors, the result is proportional to the Eigenvector (it has the same directional sense), where the constant of proportionality is the Eigenvalue (Korn and Korn, 1961).

$$\mathcal{A}\mathbf{x} = \lambda\mathbf{x}, \text{meaning} \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots a_{2,n} \\ \cdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Scheme 4.1. Two ways to write the same Eigenvalue problem.

Equation 4.1 $H\Psi = E\Psi$

What differentiates one Eigenvalue problem from another is the way that the elements of the square matrix are defined. In the Schrödinger equation describing electrons in molecules, the matrix elements are complicated integrals involving the basis functions (usually atomic orbitals) that describe the problem, and the Eigenvectors are linear combinations of the basis functions, giving a mathematical description of the molecular orbitals.

In PCA (Equation 4.2), the matrix C is referred to as the covariance matrix, while v is the set of Eigenvectors, and D is the set of the Eigenvalues. Because the goal of PCA is to find the greatest extents of variance in a set of data, the square matrix is a function of variance. Specifically, in PCA, the matrix reflects covariance. Deriving the covariance matrix C is the key to PCA, just like deriving the Hamiltonian matrix is key to solving the Schrödinger equation.

Equation 4.2 $Cv = Dv$

To generate the covariance matrix, we first take a matrix of experimental observations (*m*) for different samples (*n*) to make an *m* x *n* data matrix. For example, in array sensing, the observations may be absorbances at various wavelengths for different receptors mixed with the different analytes. The samples (number = *n*) are the individual analytes and replicates of the analytes. If we record 50 absorbance values with 5

receptors we would have 250 experimental observations ($m$ = 50 x 5) for every sample. Next, for each sample $n$, the variance in the data (experimental observations) is derived from the standard deviation, presented in Equation 4.3, where N equals the number of total observations in a group, $x_i$ is a single observation within a group and $\bar{x}_i$ is the mean of all the observations in a group. Variance is the square of the standard deviation (Equations 4.3 and 4.4).

Equation 4.3 $$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}_i\right)^2}$$

Equation 4.4 $$Var(X) = s^2$$

In our example, the data for that variance is calculated consists of all the absorbance values for the series of receptors. So far, this would mean that for each sample ($n$ of these), corresponding to potentially a large set of data ($m$ observations), we simply have one number – variance. The goal of PCA is to seek how the variance of one sample correlates with the variance of another sample. To do this, the method calculates covariance, defined as in Equation 4.5. In this formula, $x_i$ is a single observation in a group, $\bar{x}_i$ is the mean of all the observations in a group, $y_i$ is a single observation in different group and $\bar{y}_i$ is the mean of all observations in that group.

Equation 4.5 $$\mathrm{cov}(x,y)=\sum_{i=1}^{N}\frac{\left(x_i-\overline{x_i}\right)\left(y_i-\overline{y_i}\right)}{N-1}$$

Importantly, there is a covariance value for each sample relative to every other sample. Hence, for *n* samples there will be *n* x *n* covariance values. These values can therefore be arranged into the square covariance matrix (see Scheme 4.2) and this sets the stage for an Eigenvalue problem as discussed above. The matrix is symmetric across the diagonal, because the covariance of, for example sample 3 with sample 5, must be the same as between 5 and 3. In Pearson's covariance method, a specific type of PCA plot, that normalizes the data set before running the PCA algorithm, the diagonal of the covariance matrix will be equal to 1 (Molinowski, 2002).

$$C=\begin{bmatrix} cov(1,1) & cov(1,2) & \dots & cov(1,n) \\ cov(2,1) & cov(2,2) & \dots & cov(2,n) \\ \vdots & \ddots & & \vdots \\ con(n,1) & cov(n,2) & \dots & cov(n,n) \end{bmatrix}$$

Scheme 4.2: An *n x n* matrix used in PCA, where cov = covariance, and the variance of the data for each sample samples is labelled with a number up to n samples.

In the original data matrix, one can define a vector for each sample in an *m* dimensional space. After PCA, each sample consists of a vector in *n* dimensional space, where each dimension reflects decreasing extents of variance between the samples. The x, y, z… coordinates for each sample in the new space are called the scores for that sample, and the score values along each axis reflect the extent to that the samples differ along the variance expressed by that axis. The extent of variance along each principal component axis is the Eigenvalue for that axis.

The above example presents what is generally called eigenvalue value decomposition (EVD). We have used it as an example, along with MOT, in order to allow the reader to better conceptualize the underlying mechanics of PCA. However this is, in fact, only one approach to solve the Eigenvalues and is limited by requiring a square matrix. While this is a legitimate method of calculating the Eigenvalues, it is very computationally taxing A more generalized approach to the problem would be singular value decomposition (SVD). PCA can be thought of in the general form presented in Equation 4.6 where T and P' are matrices that capture the underlying data pattern of X, that is in this case the covariance matrix (Wold et al. 1987).

Equation 4.6 $$X = TP'$$

For the purposes of PCA, T is a matrix where the columns contain the factor scores for each component and the P' is a matrix where the rows contain the loading

scores. Fundamentally the factor scores are the coordinates of each sample and loading scores are the coordinates of each variable in a data reduced space. (Wall et al., 2003; Klema and Laub, 1980) SVD has the form presented in Equation 4.7. In this case the columns of V' are termed the right singular vectors and are equivalent to the columns in P. U is

Equation 4.7 $$X = UDV'$$

equivalent to T except that the length of the vectors have been normalized to 1. The diagonal elements of D are the singular values of X'X that are the square roots of the Eigenvalues of X'X. The example used by MOT solves for V' and S by diagonalizing X'X and then solving for U.

Another alternative to EVD for solving for Eigenvectors is Nonlinear Iterative Partial Least Squares (NIPALS). In this method possible loading and score values are set initially and iteratively modified until convergence between the previous values and new values is attained. (Wold et al., 31 and Risvik, 2007)

More simply, PCA rotates and combines the original data such that each new orthogonal axis explains the most possible variance. This is referred to as a change of basis. This results in the apparent shifting of the data points such that they are centered around the origin of each axis. It is here where the real strength of PCA arises. It is

generally safe to assume that the most important variables are those with the greatest variance. However it is not always apparent that combination of variables will yield a vector that explains the most variance. Furthermore, when a large data set is generated, such as for spectroscopic data, there is a fair amount of redundancy between variables. PCA fundamentally reduces the dimensionality of the data by removing redundancy and finding collinear variables and expressing them across a single axis. In essence, PCA finds the axis that best fits an n dimensional space of data and projects those axes in a simpler space (Lavine and Rayens, 2009).

The Eigenvectors of any matrix, not just with PCA, can be viewed as a "coordinate system" that is optimal for the problem under consideration. For example, if one does PCA for a set of x,y data points, the Eigenvectors correspond to two lines in the x,y plane. Along one of these lines, the variance of the data points is maximized (the data exhibits a wide range of values along that axis), while the other axis has the opposite behavior (the data exhibits a narrow range of values). While it can be quite easy to see visually what these axes are in a two-dimensional case, the generalization to more dimensions is less easily visualized, but no less straightforwardly amenable to computation.

Discriminant analysis (DA) is another Eigenvalue problem, and has many features in common with PCA. The main difference between DA and PCA is that with PCA there is no bias placed on finding the greatest variance between samples. This means that replicates of the same analyte are treated identically as different sets of analytes. Therefore, clustering of the samples in PCA means that the variance between these

samples is indeed smaller than the variance with other samples. In DA, the mathematics place a bias toward clustering repetitive samples (called a class) and separating them from repetitions of a different set of analytes (a different class).

Unlike PCA, variance is not the parameter used to distinguish data in DA. Instead, DA fundamentally finds the best way to organize data in order to maximize class discrimination. For this manuscript, the percent captured values for the PCA plots represent variance captured, and the percent captured values for the DA plots is discrimination captured. An important distinction that needs to be made is the precise form of DA used here. For the sake of simplicity, the more general form of DA called canonical discriminant analysis (CDA) will be used here. In the most basic sense, CDA identifies some combination of variables that maximize the Euclidean distance between groups while minimizing the distance between members of a group. There are other forms of discriminant analysis such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). LDA differs from CDA in that rather than relying on both within and between group data to classify data, LDA uses distance from a centroid to classify data (Lavine, and Rayens, 2009). QDA is a more complex application of the likelihood classification. However, rather than finding the linear combination of variables it identifies a quadratic surface that minimizes misclassifications (Friedman, 1989). CDA is a function that maximizes the difference between the means of differing classes, while minimizing the difference within a class. This is done by defining the scatter within a class and the scatter between the classes. Scatter is defined by matrices that ar1e analogous in form to the covariance matrices used in PCA (Fukunaga, 1990).

Importantly, two matrices are used in CDA, one for between class scatter ($S_B$) and one for within class scatter ($S_W$) (i.e. variance). Given this, the Eigenvalue problem is formulated as in Equation 4.6. The inverse of the within class scatter matrix multiplied by the between class scatter matrix acts to maximize between class scatter while minimizing within class scatter. The Eigenvectors (w) represent weighted combinations of scatter within and between the samples, while the Eigenvalues (J) represent the extent that scatter is best maximized between classes and minimized within classes. The J values are analogous to the extent of variance Eigenvalues found in PCA. Because DA has a bias built into the mathematical approach, it is called a "supervised" routine while PCA is "unsupervised". Consequently, due to the supervised nature of DA, the resulting plots often show better analyte classification than a corresponding PCA plot.

Equation 4.6 $$S_w^{-1} S_B \boldsymbol{w} = J\boldsymbol{w}$$

Now that a summary of the Schrödinger equation and PCA/DA Eigenvalue problems has been given, we can draw an analogy between the results of the two kinds of problems. The Eigenvectors of the Schrödinger equation are linear combinations of atomic orbitals we interpret as molecular orbitals. Each value of the Eigenvector is the coefficient of the atomic orbital that contributes to the molecular orbital, and it acts as a weighting factor for that atomic orbital. Each element in an Eigenvector for a particular Eigenvalue from PCA is the coordinate position of individual samples along different

axes in the *n*-dimensional space. The dimensions in this new space are orthogonal, and are referred to as PC 1, 2, …*n*, where each PC is associated with an Eigenvalue. The Eigenvalues from the Schrödinger equation are those associated with the HOMO and LUMO, meaning the orbitals near the middle of the energies. In PCA, the Eigenvalues are the extent of variance carried by each axis in the *n*-dimensional space. It is the first few principle coordinates that are the most important because they reflect the greatest amount of variance between the samples.

In a PCA or DA plot resulting from differential sensing, the response from multiple receptors can contribute to each axis in the plot, although some receptors often have a much larger contribution to a particular axis than others. The power of PCA and DA becomes most apparent in the cases that have data sets with a large number of receptors, spectral data, or other experimental data where it is nearly impossible to comprehensively evaluate the raw data with a few simple calibration curves.

It is important to note that PCA and DA are not the only algorithms used for pattern recognition. Factor analysis (FA), partial least squares (PLS), maximum redundancy analysis (MRA), and hierarchical cluster analysis (HCA) are examples of alternatives (Molinowski, 2002; Bratchell, 1987; Wold, 1973). For example, when there are many more variables than samples, DA in particular, may not perform well due to an issue called “over-fitting” which will be discussed later in this manuscript.

As already mentioned, PCA and DA are common techniques employed to analyze the data that result from differential sensing. The receptors used in this technique are commonly referred to as differential, or cross-reactive. The terms are often synonymous,

and we use them in this manner. However, for purposes of this discussion, we give them slightly different definitions. Differential receptors simply show distinct individual responses to the analytes. Cross-reactive receptors actually display trend differences in their affinities to the analytes, meaning that some receptors have higher affinities to some analytes, while the corresponding cross-reactive receptors prefer different analytes. This means that cross-reactive receptors are a subset of differential receptors. Finally, given these definitions, highly selective receptors are clearly both cross-reactive and differential.

## 3. Model Setup

In order to illustrate the use of PCA and DA in differential sensing we will present a variety of artificial data sets that show behavior similar to what one might see in an array-sensing experiment. These data sets were created by the authors to display a few typical behaviors observed in these types of analysis. The analyses in this manuscript are model examples, carried out to illuminate certain points. A set of five hosts (receptors) and five guests (analytes or mixtures) a were generated where each host-guest measurement is modeled as if it was repeated five times. The measurements are host:guest binding constants ($K_a$ values), although they could represent any kind of data such as spectral intensities. However, by using $K_a$ values the discussion naturally has lessons related to the selectivity of the receptors.

For each scenario, values were selected to represent the $K_a$ of each host:guest pair. For each pair, five values representing repetitions were randomly generated, following a

normal distribution (Samuels and Witmer, 2003). This distribution was set such that the mean of the values was equal to the $K_a$ value selected to represent the host:guest pair. The 0.5 to 5 standard deviations ($\sigma$) of the distributions of $K_a$ values for each host:guest pair were used in order to simulate a range of variances within repetitions. For each scenario presented we have included a summary of the mean $K_a$ values used and the $\sigma$-value used.

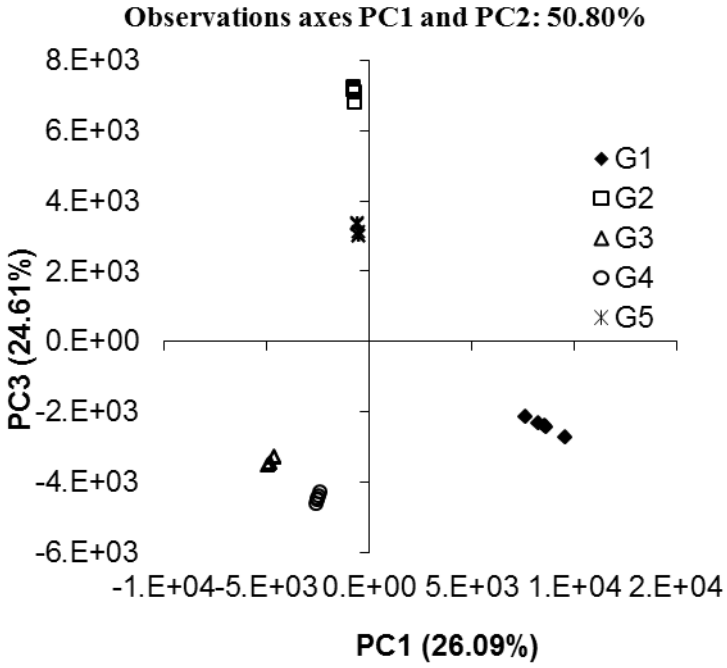
# 4. Exploration of PCA and DA

## 4.1 Lock-and-Key Array versus Cross-Reactive Array

Most times when examining DA and PCA score plots, receptor performance may not be entirely known. However, careful examination of the plot results can shed some light on how a receptor is performing. Consider a situation in that there is a panel of antibody-like receptors that are highly selective, each to different individual guests (Figure 4.1) and a very low $\sigma$ relative to the $K_a$ values. The resulting score plot shows that each guest occupies a distinct location in the PCA plot. G1, G3, and G4 are discriminated primarily across the F1 axis (principal component 1, PC1) while G2, and G5 are discriminated primarily across the F3 axis (principal component 3, PC3). (Note that F1 was plotted versus F3. This was chosen in order to better display the visual separation of each of the guests. We will be discussing methods and rationale for improving visual discrimination later in this PC1) It is important to note that approximately 50% of the variance is found in the F2, F4, and F5 axes. In this particular case all of the guests can be visually

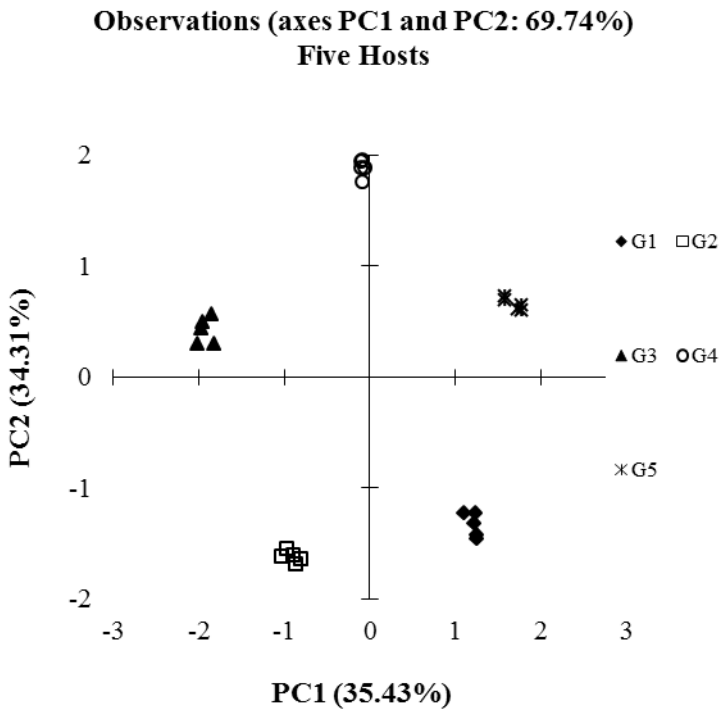
discriminated by two axes, though there is still significant discrimination in the remaining axes.

Similar results can be achieved through lower selectivity but fully cross-reactive receptors (Figure 4.2). Figure 4.2 represents a plot in that each receptor is cross-reactive with all other receptors. In this case, because each of the host:guest pairs behaves in a unique manner each guest is separated from the others in both the F1 and F2 axes. From this example, it seems that there is very little difference in using a panel of receptors that have antibody-like behavior as opposed to cross-reactive behavior, since discrimination of analytes can be effectively achieved in both circumstances. In these models, each host responds in an unambiguously different manner to all the guests. This situation is ideal for optimal discrimination. However, quality discrimination can still be achieved with small differences between responses to guests for an array of hosts, as is the case in most cross-reactive arrays, since each receptor behaves in a sufficiently unique manner. These preliminary conclusions support the notion for utilizing cross-reactive arrays, that generally require far less synthetic effort to develop than antibody-like highly selective receptors.




| | H1 | H2 | H3 | H4 | H5 |
|---|---|---|---|---|---|
| G1 | 1000 | 10 | 10 | 10 | 10 |
| G2 | 10 | 1000 | 10 | 10 | 10 |
| G3 | 10 | 10 | 1000 | 10 | 10 |
| G4 | 10 | 10 | 10 | 1000 | 10 |
| G5 | 10 | 10 | 10 | 10 | 1000 |

Figure 4.1. A) PCA plot of the antibody-like scenario and mean $K_a$ values for the "antibody like" scenario. In this example, each host:guest behaves in a very specific manner. For example, Guest 1 (G1) and Host 1 (H1) have a very high affinity for each other relative to the other host:guest pairs (0.5 standard deviations).

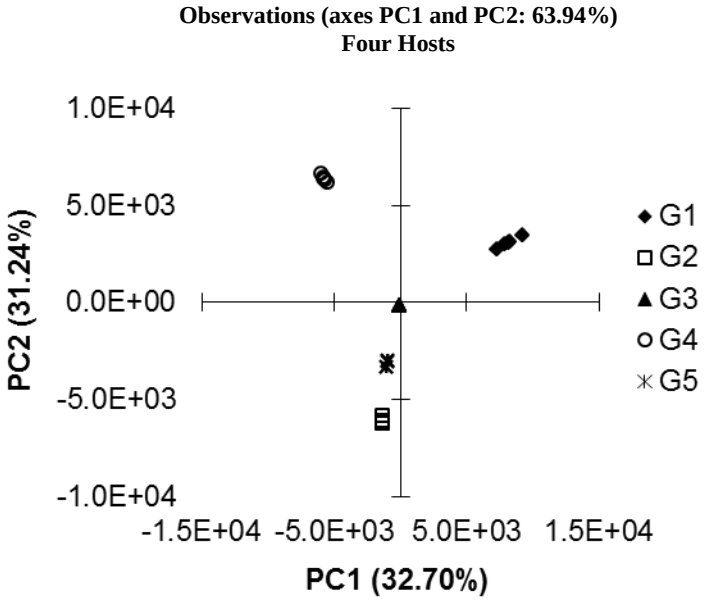



| | | | | | |
|---|---|---|---|---|---|
| G1 | 2.5 | 10 | 18 | 13 | 8 |
| G2 | 8 | 2.5 | 22 | 18 | 13 |
| G3 | 13 | 10 | 2.5 | 22 | 18 |
| G4 | 18 | 10 | 8 | 2.5 | 22 |
| G5 | 22 | 10 | 13 | 8 | 2.5 |

Figure 4.2. PCA plot of the cross-reactive scenario and mean $K_a$ values for the cross-reactive scenario. In this example, each host:guest behaves in a very unique manner. For example, Guest 1 (G1) and Host 1 (H1) have lower affinity for each other than the affinity of H1 for any of the other guests, whereas Host 2 (H2) has the lowest affinity for G2 relative to the other host:guest pairs (2 standard deviations).

## 4.2 Choosing the Best Number of Hosts for an Array

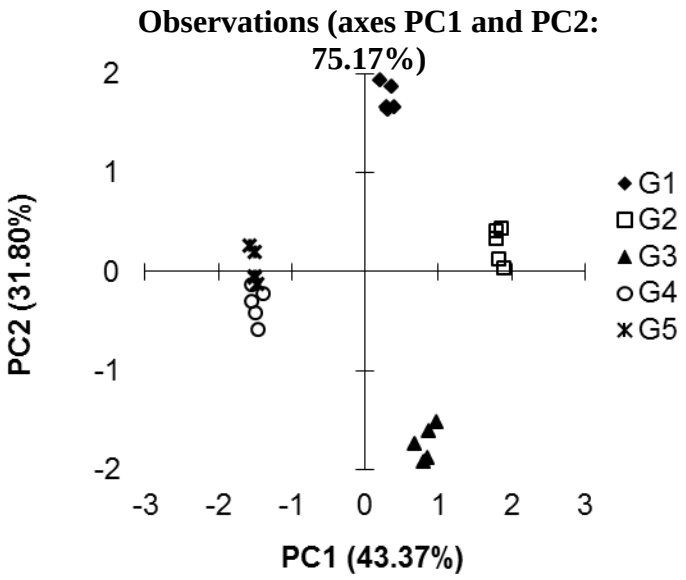
Care must be taken in selecting the correct number of hosts, be it high selectivity or cross-reactive array. Studying the number of receptors needed for discrimination may be helpful in such a case. Figure 4.3 portrays the same data as Figure 4.1 but using four hosts instead of five. This is an example where exploring the number of receptors used explain the similarities between the two figures. The analysis reveals that a lack of a signal can be just as important as a measurable single in an array setting. In Figure 4.3, guest 3 does not respond to any of the hosts. Its behavior is therefore different than the other guests and can easily be separated from the remaining guests. Essentially, in arrays where there is high selectivity, fewer hosts can be used to achieve optimal discrimination in many cases. This can be thought of in terms of a combination of 1s and 0s. An antibody-like sensor can be considered "perfect," when presented with its target, it has maximum signal and can be assigned a 1. When presented with a non-target analyte it has no signal and can be assigned a 0. A combination of 3 receptors could have the values (1,1,1), (1,1,0), (1,0,1), (0,1,1), (1,0,0), (0,1,0), (0,0,1) or (0,0,0) giving eight unique combination. However the number of receptors needed to generate all these unique patterns is only 3 not 8. Cross-reactive arrays are not constrained by 1 and 0 values. Rather, they are limited only by their ability to create a reproducible and sufficiently unique pattern of binding for each target.




| | H1 | H2 | H4 | H5 |
|---|---|---|---|---|
| G1 | 1000 | 10 | 10 | 10 |
| G2 | 10 | 1000 | 10 | 10 |
| G3 | 10 | 10 | 10 | 10 |
| G4 | 10 | 10 | 1000 | 10 |
| G5 | 10 | 10 | 10 | 1000 |

Figure 4.3. PCA plot of antibody-like scenario with four hosts and mean $K_a$ values. This is the same data sets as presented in Figure 4.1, however one of the hosts has been omitted (0.5 standard deviations).

When a single host is removed from the cross-reactive data set in Figure 4.2, the score plot still retains a high level of discrimination (Figure 4.4). However, Guests 2 and 4 begin to show overlap in a two-dimensional plot. This is due to the lack of a sufficient number of hosts behaving in a distinctly unique manner. Also, in the cross-reactive case an increase can be seen in the overall variance of PC1 and PC2, from 69.74% to 75.17%. This is due to the properties of the data set itself. In x,y plots of multivariate data, there is variance found in each of the variables but only two dimensions are displayed. Therefore, in a five variable array, the variance is distributed across five dimensions. When a variable is removed from the system, there are fewer dimensions across that the variance can be distributed.

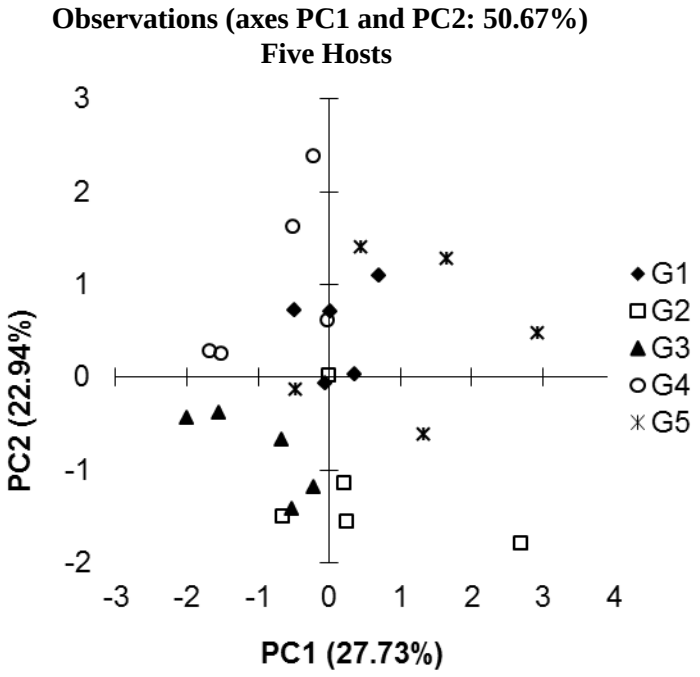



| | H1 | H2 | H3 | H4 |
|---|---|---|---|---|
| G1 | 2.5 | 22 | 18 | 13 |
| G2 | 8 | 2.5 | 22 | 18 |
| G3 | 13 | 8 | 2.5 | 22 |
| G4 | 18 | 13 | 8 | 2.5 |
| G5 | 22 | 18 | 13 | 8 |

Figure 4.4. PCA plot of cross- reactive scenario with four hosts and mean Ka values. This is the same data set as presented in Figure 4.2, however, one of the hosts has been omitted (2 standard deviations).

### 4.3 When to Add Hosts to an Array

An array where every host:guest pair generates a signal only marginally distinguishable from background noise occurs when different host:guest pair shows very similar affinities to analytes, yet the variance in the signal results in some overlap between signal derived from back grand and signal derived from the specific target. In Figure 4.5 each host has only one “best” guest but there is significant overlap between the “best” values and the non-specific interactions. This can be considered analogous to the antibody-like scenario, except with much lower selectivity. The following section will explore a similar situation found in cross-reactive systems where the affinity for each host:guest pair varies only slightly. In Figure 4.5 clustering of the various guests may roughly exist, but the groups are not readily distinguishable. This is due to a high standard deviation between the repetitions, relative to the magnitude of the $K_a$ values for all the guest groups. Figure 4.5 is the PCA plot of the data, that appears as total scatter. However, even with DA (Figure 4.6), the method falls short of completely discriminating the analyte classes. One possible reason for this lack of discrimination has to do with what characteristics are being used to classify the data.




| | H1 | H2 | H3 | H4 | H5 |
|---|---|---|---|---|---|
| G1 | 10 | 8 | 8 | 8 | 10 |
| G2 | 8 | 10 | 8 | 8 | 8 |
| G3 | 8 | 8 | 10 | 8 | 8 |
| G4 | 8 | 8 | 8 | 10 | 8 |
| G5 | 8 | 8 | 8 | 8 | 10 |

Figure 4.5. A) PCA plot of overlapping data set with 5 hosts and mean $K_a$ values (x 100) for each host guest pair (5 standard deviations).

As is typical with PCA, it is assumed that variance between groups of guests is sufficient to categorize data. DA on the other hand requires information about class membership to group the data. This situation is hinted at by the difference in variance

captured by the PCA in Figure 4.5 and the percent of discrimination captured by the DA in Figure 4.6. If the amount of variance captured by the first few PCs in PCA is

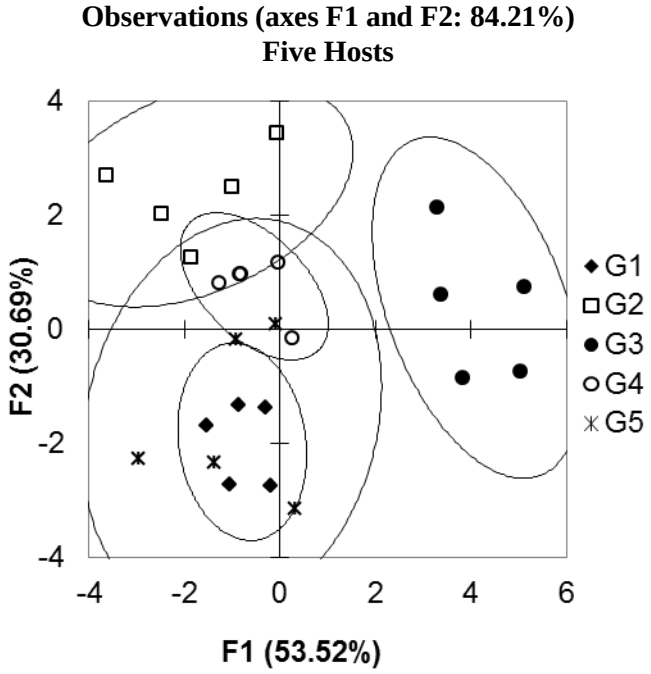



Figure 4.6. DA plot overlapping data set with 5 hosts.

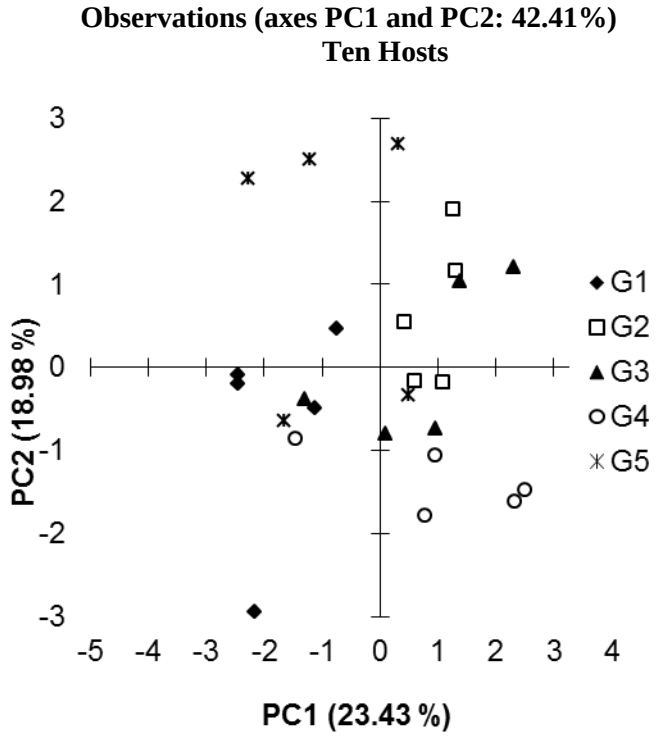
low then it is possible that variance is not a good classifier. However, this leads to the debate of how many components are appropriate to include in a model. There is some sentiment a model with many components each capturing a small amount of variance is better that a model with few components each capturing amlarge amount of variance (Janzen, 2006). The counter point to this is that by including many components in the analysis, PCA's main objective to reduce the dimensionality is neglected. To some extent, this is a philosophical question we will not be exploring here. However, careful

consideration of what exactly a researcher wishes to achieve as well as how the data is expected to behave may allow one to decide what the best model approach may be.

While the DA appears to be superior to the PCA in 4.5 due the increased amount of discrimination captured, it likely will perform poorly as a predictive model. This is due to the model relying on all data points in order to create the classifier model. A leave-one-out cross validation excludes each data point iteratively and the analysis is performed without the omitted point. Each new model is then used to predict the class of the excluded point. When this sort of analysis is performed on a data set where the magnitude of the variance is such that each sample is essential to discrimination one can see that the model loses its predictive power.

In cases like this, adding additional hosts can improve the data discrimination by reinforcing patterns in the data sets that are difficult to observe. Figures 4.7 and 4.8 show the PCA and DA plots obtained when ten hosts are considered, rather than the five used for Figures 4.5 and 4.6. The five new hosts were chosen to respond identically to the first five. This could represent additional replicates in the system, or additional hosts where the deviation from other hosts is subtle. In Figure 4.7 we see that each guest is more localized in the PCA plot, though overlap still exists. In the DA plot (Figure 4.8), there exists a much tighter clustering of the guests. The improvement can be further supported by considering the jack-knife analysis for the five host data set DA plot (76%), and the improved jack-knife analysis for the ten host data set DA plot (84%).

The reason additional hosts improved the discriminatory power of this system is that each host responds to the guests in a specific manner that is not easily observed due




| | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 |
|---|---|---|---|---|---|---|---|---|---|---|
| G1 | 10 | 8 | 8 | 8 | 10 | 8 | 8 | 8 | 8 | 8 |
| G2 | 8 | 10 | 8 | 8 | 8 | 10 | 8 | 8 | 8 | 8 |
| G3 | 8 | 8 | 10 | 8 | 8 | 8 | 10 | 8 | 8 | 8 |
| G4 | 8 | 8 | 8 | 10 | 8 | 8 | 8 | 10 | 8 | 8 |
| G5 | 8 | 8 | 8 | 8 | 10 | 8 | 8 | 8 | 10 | 8 |

Figure 4.7. PCA plot of overlapping data set with ten hosts.

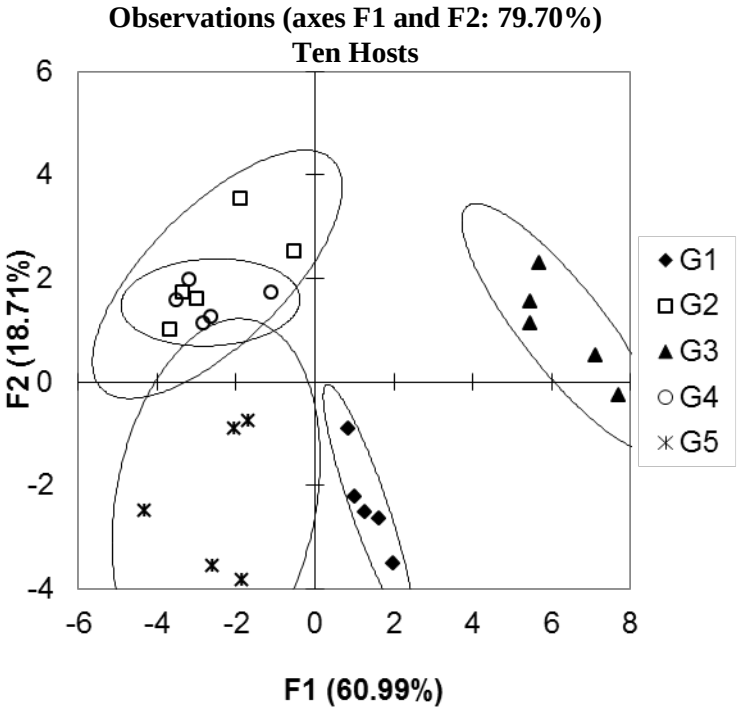



Figure 4.8. LDA plot of over lapping data set with ten hosts.

to a high amount of noise in the system. In such circumstances, differentiation of analytes can be achieved by adding additional hosts that either reinforce observed patterns by adding additional hosts with similar behavior (as the example presented here does), or by adding additional hosts with wholly unique behaviour. This situation, where adding hosts to a high noise system increases the discriminatory power of an array, is called co-linearity (Bajaj at al., 2010).

When this data set is expanded to 20 hosts (Figure 4.9) - a situation is created were there are many more hosts than guests and the amount of variance captured across the first few

PCs decreases. This is not unexpected, as each PC captures a portion of the variance; as the number of PCs increases the amount of variance captured by each PC decreases. Visual examination of the first 2 PCs appears to show superior grouping of the samples. However when a validation method is applied is it is found that the

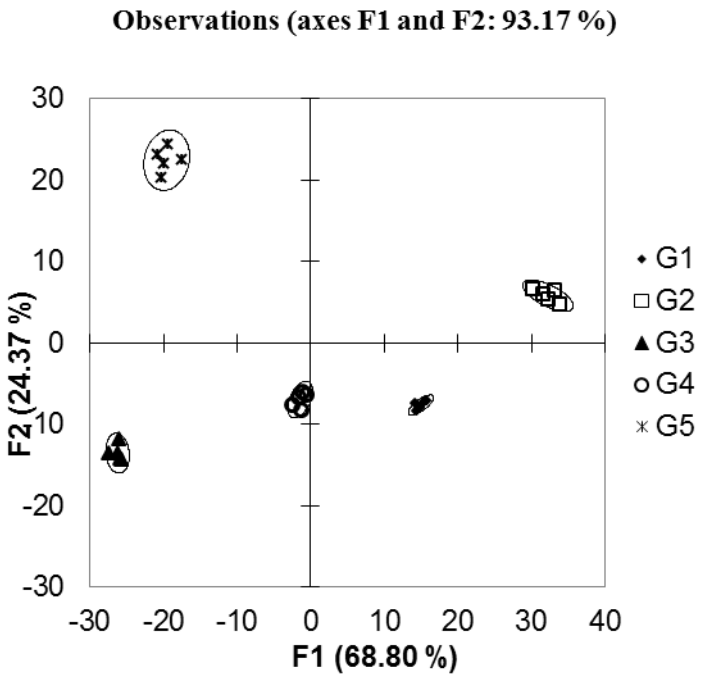



Figure 4.9. PCA of overlapping data with 20 hosts. The data has been "over-fitted".

model has very poor predictive power, in this case the jack-knife analysis yields a 12% correct classification rate.

This is a situation referred to as over-fitting and is a common trap many researchers fall into. Even in the most random data set, it is possible to find an equation

that can perfectly group the data by whatever parameter the researcher wants. If too much data is used, however, the equation is only relevant to the presently available data. Any new data is not likely to follow the lines of discrimination, resulting in a model that only predicts itself. When examining the quality of a PCA or DA one must be cognizant that adding more variables can appear to make a better fitting model while reducing the predictive power of the model (Tobias, 1995).

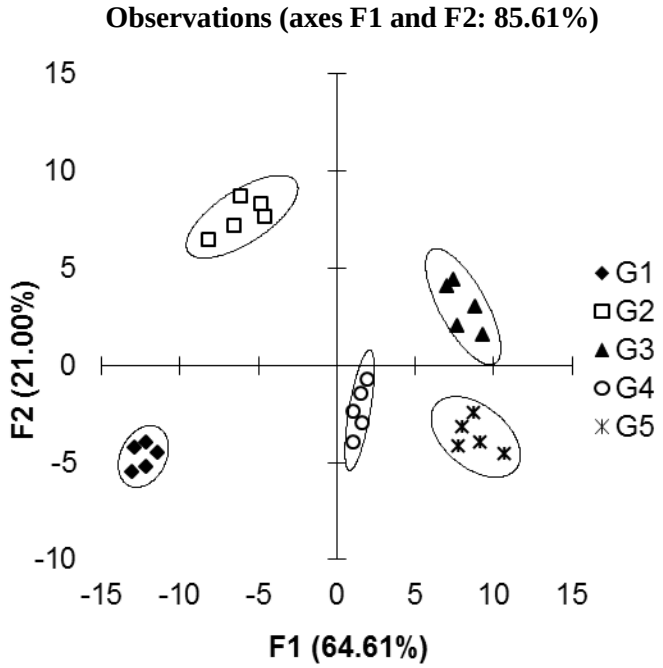
### 4.4 High Dimensionality in an Array and Determining Host Performance

High dimensionality in PCA and DA plots is often a goal for many authors during the statistical analysis of their data (Palacioet al., 2007; Zhang and Suslick, 2007; Suslick, 2004; Suslick and Rakow, 2004). High dimensionality is defined in PCA and DA as a large number of principal components or discriminating axes, all of that carry a significant extent of the total differentiation. High dimensionality is desirable in cases where very similar analytes need to be differentiated. However, in many circumstances where high dimensionality exists, a two or three-dimensional plot, while mathematically aiding in the differentiation of analytes, may not lead to an optimal pictorial representation of discriminated data. High dimensionality, and thus more principal components or discriminating axes, is obtained by adding more cross-reactive or highly selective receptors to an array. This makes sense, as the number of discriminating axes possible in a PCA or DA plot is directly correlated to the number of receptor variables.

The math behind the decomposition of a data set in PCA into its corresponding Eigenvectors necessitates that the number of Eigenvectors that emerge from the

calculation be equal to the number of receptors in the array or the number of samples, whichever of these two numbers is smaller. In most circumstances of differential sensing, the number of receptors in an array will be smaller than the number of analytes, and thus the number of Eigenvectors that arise in PCA will be equal to the number of receptors (Molinowski, 2002). Similarly, the number of Eigenvectors that can emerge from a DA is equal to the number of classes minus one (Johnson, and Wichern, 1992). However, this correlation between the number of Eigenvectors and number of receptors often leads to an incorrect conclusion: Each discriminating axis represents one receptor (variable). To further understand why this conclusion is a misconception, we must turn to loading plots, which are simultaneously generated when PCA and DA plots are produced. Loading plots show the influence that each receptor, or variable, has on the corresponding discriminating axis. Each receptor is represented by a vector in a loading plot. The x,y coordinates (or higher coordinates) of each vector indicate the extent to that each receptor contributes to a discriminating axis. Vectors of (1,0) or (1,0) most influence the discrimination of analytes along the x-axis (F1), with the vector (-1,0) best discriminating analytes on the left side of F1 and vector (1,0) best discriminating analytes on the right side of F1. Conversely, vectors (0,-1) or (0,1) most influence the lower half of the y-axis or the upper half of the y-axis, respectively. Receptors with vectors of intermediary x,y values indicate contributions to both axes. Thus, the loading plot is used to explore which receptors or variables are most useful for discrimination, thus aiding in determining receptor performance.

Similar to a loading plot is a biplot. Biplots are most commonly seen in conjunction with PCA plots, where the loading plot is superimposed onto its corresponding PCA plot. In these plots, the receptors that most influence a particular data point are located close in vector space to the data point. The proximity of a receptor vector endpoint to a data point allows further analysis of the array system to determine whether the receptor is important for discriminating that particular analyte.The loading plot and biplots make it clear that the differing principal components can be made up from several responses of receptors. Both loading plots and biplots are important plots to analyze once PCA or DA results have been generated. They allow the user to probe the importance of the receptors in an array, that in turn provides information to improve and modify the array to obtain the best results. The DA plot in Figure 4.10 was derived from an array consisting of 15 receptors, and the mean $K_a$ values used for this simulation. These values were chosen to maximize the dissimilarity between the behavior of each host:guest interaction in order to observe how each receptor can contribute to multiple axes. In the loading plot (Figure 4.11), we see that the first axis (F1) discriminates analytes based on the




| | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 | H15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G1 | 3 | 23 | 18 | 13 | 8 | 3 | 13 | 13 | 8 | 23 | 23 | 23 | 18 | 3 | 3 |
| G2 | 8 | 3 | 23 | 18 | 13 | 13 | 18 | 8 | 3 | 13 | 13 | 8 | 3 | 8 | 8 |
| G3 | 13 | 8 | 3 | 23 | 18 | 18 | 8 | 3 | 13 | 3 | 18 | 28 | 23 | 23 | 18 |
| G4 | 18 | 13 | 8 | 3 | 23 | 23 | 23 | 18 | 23 | 8 | 8 | 13 | 13 | 13 | 13 |
| G5 | 23 | 18 | 13 | 8 | 3 | 8 | 3 | 13 | 18 | 18 | 3 | 3 | 8 | 18 | 23 |

Figure 4.10. DA plot of 15 hosts with co-linear variable and high variance and the mean $K_a$ values for unique host behavior data set (2 standard deviations).

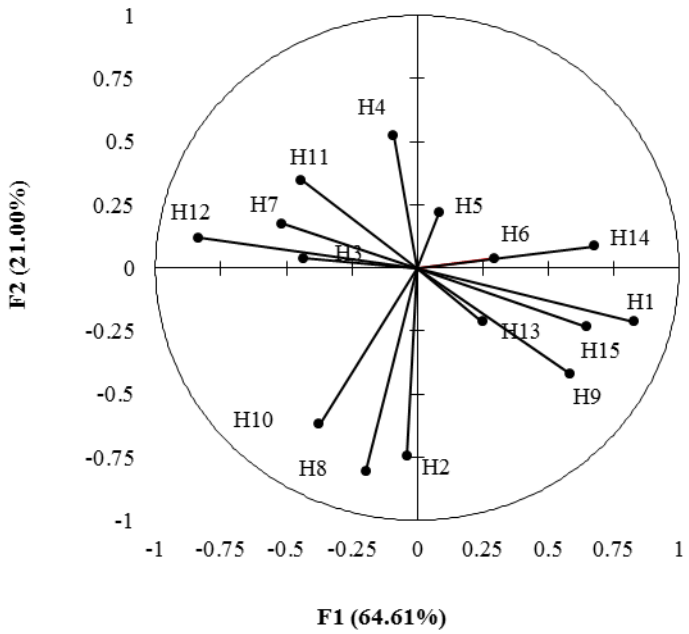



Figure 4.11. Loading plot of the DA plot in 6A, identifying the contribution of each host to an axis.

data from all the receptors, except H2 that shows a value close to zero as an x-component in its line ending vector. The second axis (F2) discriminates analytes based on the data from nearly all the receptors, because most receptors contain a non-zero value as a y-component in their line ending vector (H3 and H6 are near zero). Receptors H2, H8, H10, and H4 contribute the most to the discrimination seen with the F2 axis because the absolute value of their y-component is larger than the absolute value of the y-vector components of the other receptors. Therefore, each discriminating axis contains contributions from multiple variables.

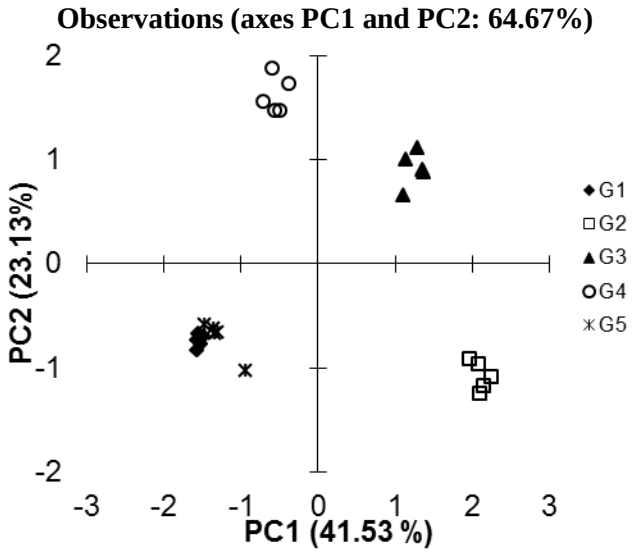
Care must be taken not to rely on loading plots and biplots exclusively for receptor selection. While it is likely that H2 and H8 are the primary contributors to the position of G4 in Figure 4.10, the precise relationships between the host's influence on the guest's position cannot be determined. Biplots are useful tools for approximating the host's influence however, for a more exact measure and for optimal variable selection A factor analysis could be performed. This method uses the loadings and biplots and seen in PCA, and applies a set of rules and criterions in order to quantify the relative significance each factor has on the data structure of the model (Molinowski, 2002). This gives a quantitative estimation of the importance of each factor. However, in many cases PCA will yield equivalent results to a factor analysis (Fabrigar, 1999).

The misconception that each discriminating axis represents one receptor in an array is most likely a result of the direct correlation seen between the number of receptors (variables) in an array and the number of discriminating axes obtained. Another reason for this misunderstanding may be that in high dimensionality systems with a large number of receptors, it is often the case that only a few of the receptors have pertinent contribution to a particular discriminating axis, while the other variables in the array have a very small contribution that can be considered negligible.

## 4.5 Obtaining the Best Visually Representative Plot

After running PCA or DA algorithms, many statistical programs automatically generate a two-dimensional plot using the two discriminating axes that contain the maximum variance or discrimination. Oftentimes, this leads to a satisfactory plot,

however, there are circumstances where this automatically generated two-dimensional PCA or DA plot may not be the best visual representation of the data. In cases such as this, it becomes important to consider all of the components computed by the statistical analysis program. Figure 4.12 shows an example where a two-dimensional




| | H1 | H2 | H3 | H4 | H5 |
|---|---|---|---|---|---|
| G1 | 10 | 10 | 10 | 10 | 10 |
| G2 | 9000 | 800 | 850 | 650 | 1000 |
| G3 | 8500 | 700 | 800 | 550 | 750 |
| G4 | 7000 | 500 | 750 | 450 | 650 |
| G5 | 6000 | 400 | 700 | 350 | 400 |

Figure 4.12. PCA plot with a large variance data set and the mean $K_a$ values of inconsistent variance data (Data for G1 contains 0..5 standard deviations, data for G2-G5 with $K_a$ values of 10 contains 0.5 standard deviations, data for G2-G5 with $K_a$ values of 20 contains 1 standard deviation, data for G2-G5 with $K_a$ values of 30 contains 1.5 standard deviations, data for G2-G5 with $K_a$ values of 40 contains 2 standard deviations).

plot of the data shows considerable overlap between the analytes G1 and G5. The $K_a$ and σ values employed in this example (presented in Figure 4.12) were chosen to generate this coincidental overlap.

Once we examine the three-dimensional plot, that takes into account a third discriminating component (Figure 4.13), we see excellent discrimination of all the

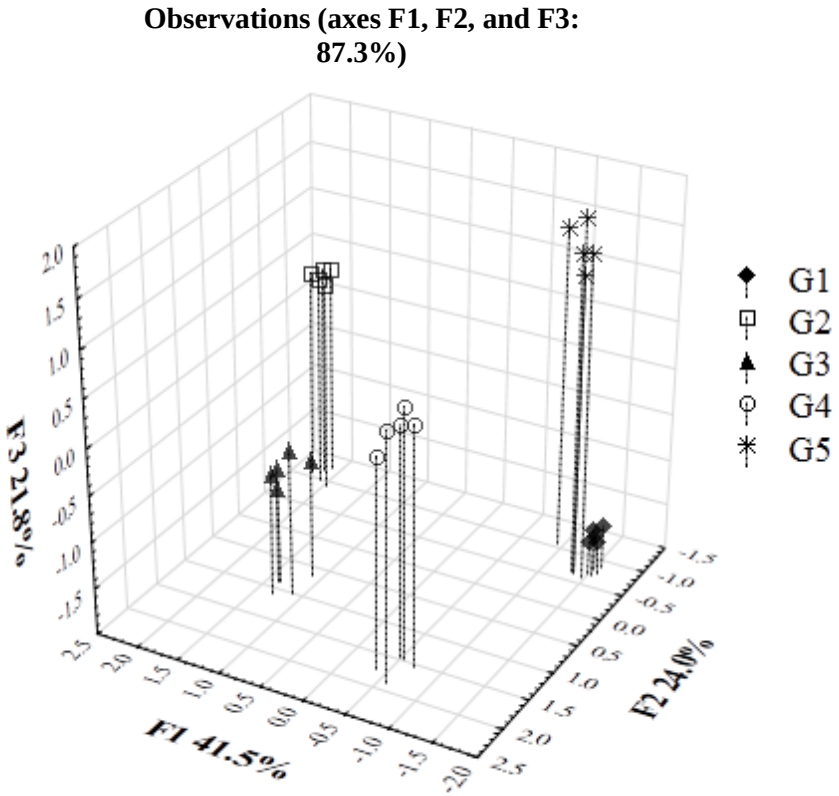



Figure 4.13. Three-dimensional PCA plot of the data set.

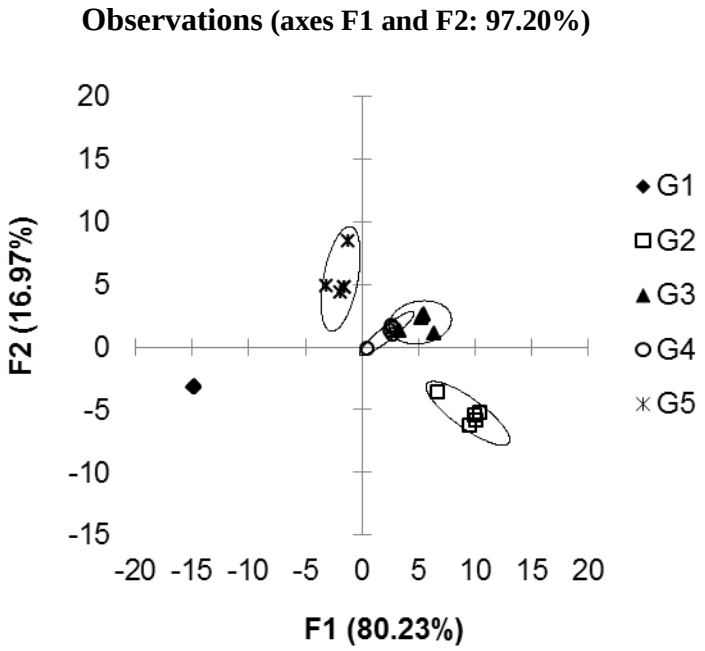
analytes. If any of the other discriminating axes calculated by the PCA or DA algorithms hold a substantial discriminating percentage (i.e. >5%), it would be beneficial to further examine those axes in addition to the two greatest discriminatory axes. For instance, there may be circumstances where a third and fourth discriminating components are important to differentiate a data set. In this case, an examination of the plots generated from all combinations of the first through fourth axes may be necessary (i.e. compare plots – 1vs.2, 1vs.3, 1vs.4, 2vs.3, 2vs.4). Thus, careful consideration of all components may lead to the best visual representation of differentiated data.

To specifically optimize PCA plots, there exist rotation methods that often aid in simplifying the discriminating axes for easier interpretation. Although there are several methods, varimax rotation introduced by Kaiser in 1958 is the most common. Varimax works by searching for a rotation of the original discriminating axes that maximizes the variance of the squared loading scores (Kaiser, 1958). The advantage of utilizing varimax is that the new plot may be easier to analyse because each axis represents a response from one receptor, or only a few receptors (Abdi, 2003). This tends to lead to loading scores (i.e. placements within the PCA plot) that have a wide range of values and emphasize clustering (Molinowski, 2002**)**.

### 4.6 Including Blank or Control Responses in an Array

Another factor that should be considered when selecting data to be discriminated by PCA or DA is whether to include blank or control responses in the data set. Researchers are often eager to show the excellent response that their receptor array shows towards the

desired analytes versus a blank or control sample. However, inclusion of blank or control data within the data set evaluated by PCA or DA can lead to artificial discrimination in the plot. The generated plot becomes skewed from the blank or control sample data. Take the following example (Figure 4.14), where G1 represents blank or control samples, which do not respond to the receptor array. In this example, the F1 discrimination is dominated by the response difference between the blank/control samples and the analytes being tested. Here, a large fraction of the F2 axis instead becomes the differentiation that is seen within the analytes tested. Once we remove the blank/control from our data set, the plot shown in Figure 4.15 is generated. It is not necessarily incorrect to include the blank or control in the data set being evaluated by PCA or DA. If the researcher's primary goal is to differentiate non-responsive samples (i.e. blanks or controls) from response samples, then including the blank/controls in the data set is appropriate. However, usually the goal is to differentiate analytes, and thus, an omission of the blank/control samples from the PCA or DA data set is generally most sensible.




| | H1 | H2 | H3 | H4 | H5 |
|---|---|---|---|---|---|
| G1 | 10 | 10 | 10 | 10 | 10 |
| G2 | 9000 | 800 | 850 | 650 | 1000 |
| G3 | 8500 | 700 | 800 | 550 | 750 |
| G4 | 7000 | 500 | 750 | 450 | 650 |
| G5 | 6000 | 400 | 700 | 350 | 400 |

Figure 4.14. A) DA plot of low variance data with blank included in the data set. And the mean $K_a$ values of low variance data (0.5 standard deviations).

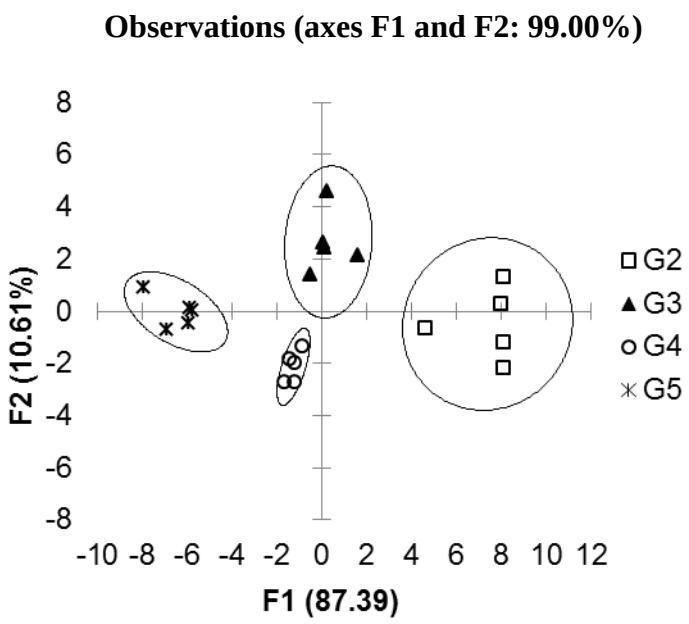
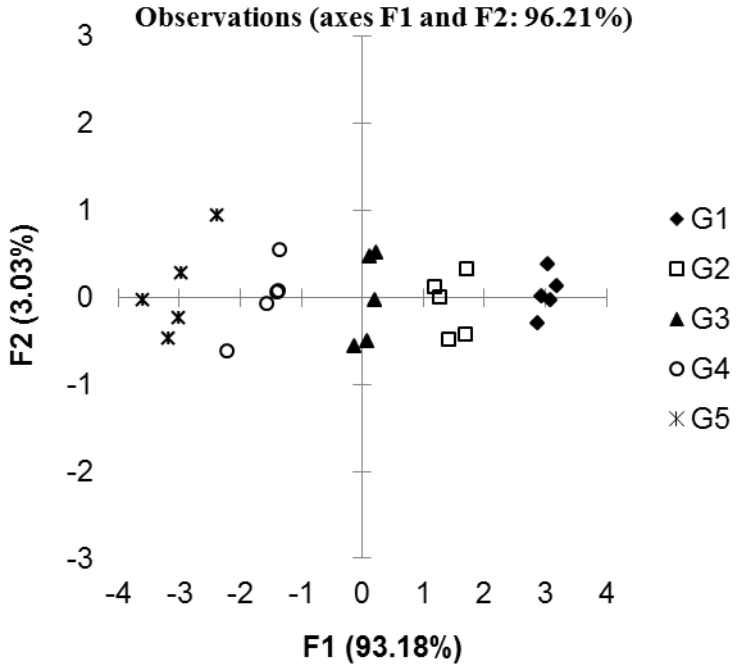



Figure 4.15. DA plot with blank excluded from data set.

### 4.7 Circumstances When Arrays May Not Be Necessary

PCA and DA find their greatest utility in circumstances where data is obtained from cross-reactive receptors. However, in cases where high cross-reactivity is not seen within receptors, repetitive data that does not assist in the differentiation of the analytes may exist. In these types of circumstances, an array of receptors may not be necessary. Instead, one single receptor is sufficient for analyte discrimination. Take for example a case where there are several receptors that have nearly identical signal response trends, with only slight differences between receptors in their intensity of overall response to the set of analytes (Figure 4.16). We see that the corresponding PCA and DA plots (Figures 4.16 and 4.17) for this example show the data tightly




| | H1 | H2 | H3 | H4 | H5 |
|---|---|---|---|---|---|
| G1 | 9500 | 10000 | 9000 | 8000 | 11000 |
| G2 | 9000 | 8500 | 8500 | 6000 | 10000 |
| G3 | 8500 | 7500 | 8000 | 5000 | 7500 |
| G4 | 7000 | 4500 | 7500 | 4500 | 6000 |
| G5 | 5500 | 4000 | 7000 | 3500 | 4000 |

Figure 4.16. A PCA plot of data where an array is not needed and the mean $K_a$ values for a plot where an array is not needed (0.5 standard deviations).

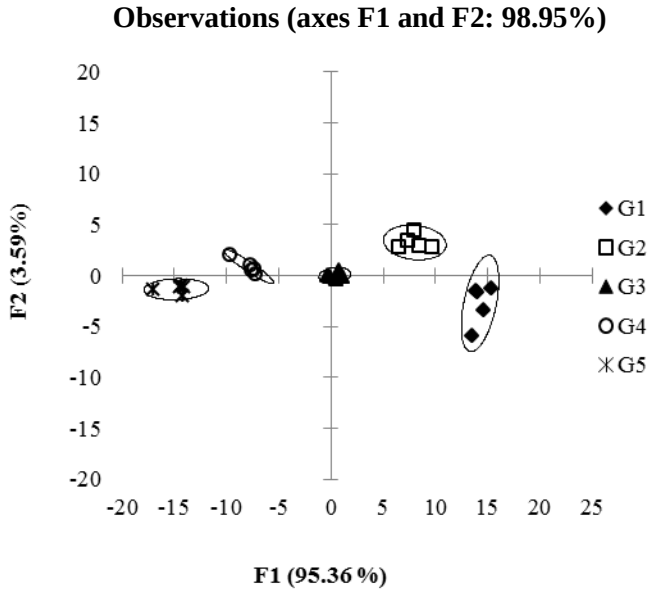



Figure 4.17 An LDA plot of data where an array is not needed.

clustered according to analyte identity along the F1 axis. Variance of the F2 axis, however, is misleading as the variance arises solely from the noise found within analyte groupings for the array. In this case, both of these plots would be better represented in a two dimensional graph (analyte vs. signal) plotted from the data obtained with only one of the receptors from the original array. Thus, the use of an array and corresponding multivariate statistical analysis tools are both unnecessary here. In circumstances like this, one receptor is sufficient for the purpose of analyte differentiation. To prevent such a case where unnecessary work and time have been spent planning, engineering, and executing an array when only one receptor from the array is needed to accomplish the

desired goal of analyte differentiation, it is prudent to always be observant for arrays in that receptors give highly similar response trends.

# 5 Practical Application of PCA and DA

## 5.1 Using PCA and DA Together as Validation Techniques

PCA and DA are both methods that are best used in concert to optimize data analysis. Typically, PCA is run first to assist in uncovering general trends in the data set. Once PCA has been run, a DA is run to specifically investigate the classification and grouping trends present in the data set. Although sometimes the graphs obtained for a data set with PCA and DA look similar, occasionally the two methods can identify different patterns in the data set. For this reason, we generally recommend using both of these methods to simultaneously explore the outcomes of these analyses for trends.

As already mentioned, it is often very common for a DA plot to appear to have better discrimination than its corresponding PCA plot. Therefore, as we have mentioned a few times, validation techniques are simultaneously run with trained models to allow users to evaluate the validity of the model for their data set. A common initial validation technique for DA is the jack-knife analysis, also known as the "leave-one-out" analysis. In this validation technique, data for one or several samples are removed from the data set and a new model is constructed. The classification of this removed analyte is then estimated from the new model and compared to its previous classification. This entire

process is completed with every analyte in the data set, and the resulting number is the percentage of classifications that were correctly identified (Molinowski, 2002).

In addition to validation techniques for DA, most computer software programs give an option to display confidence ellipses for the grouped data. These ellipses generally represent a 95% confidence limit for a specific analyte group typically calculated using the Hotelling T2 statistic. These confidence ellipses help the user to more easily identify how close each sample is to the group centroid. On this note, however, we strongly discourage the incorporation of arbitrarily drawn circles that encompass analyte groups, as these may be mistaken for confidence ellipses.

A more relevant statistic for assessing the quality of classification would be the use of bootstrapping in order to approximate the true characteristics of the population. Briefly, a bootstrap method resamples data from a sample population in order to create an expected distribution of the data. This process is repeated, often several thousands of times, until a reasonably accurate distribution is generated. This allows the researcher to generate confidence intervals that are directly related to the true distribution of the data rather than make assumptions when using the T2 statistic (Wehrens et al., 2000).

These techniques can be used to validate PCA data as well as DA data. While not considered a classification method independent, PCA scores can be used for classification. Most software packages should support validations methods for both DA and PCA though the exact title may vary. One software package may use the term “predictive PCA” (Eriksson et al., 2006) while another may call it “principal component regression”.

If the final goal of these methods is to classify unknown data, an external validation data set should be used. This is data for that the researcher knows the class (though if possible, blinding the experimenter should be considered) but was not used to develop the model. This can include data that was collected as part of an experiment but not used in the model, or data collected as part of another experiment. There are a variety of methods used to quantify the prediction power of a model using an external data set. The precise method used should be selected to best reflect how the model is expected to be used and the availability and or quality of external data (Consoni, et al., 2010)

## 5.2 Pre-processing Data

Pre-processing data is often an important step to take into consideration before running multivariate analysis methods. Frequently data is preprocessed by transformation in order to achieve certain data structures. Linearity, for example is a requirement for many models; log transformation can be used to bring data into this shape. When using attributes of the data, such as variance, to identify differences in the data, it is often prudent to normalize the data to eliminate variation that does not contribute to classification. A common method to achieve this is called centering. This involves setting the mean of each variable to some constant number. Frequently this value is zero, however it can also be some value that has meaning for the data set. Typically, scaling is used in conjunction with centering to further normalize the variables to each other. With scaling, variables with large values are fundamentally “shrunk” while variables with small variables are “stretched” to put them on the same footing (Eriksson et al., 2006).

Without these steps it is often the case that the primary vector of variance is defined by the mean of the data. This limits the utility of the model by obscuring latent variables that are more powerful for discrimination (Han et al., 2011).

Another method commonly used to remove noise from data is referred to as smoothing. One of the simplest examples of this is the moving average. In this example a window of points is averaged together to generate a single point in their place. The window is then shifted through the entire data set point by point until each new point represents an average of a subset of points. This method reduces the impact of especially high or low values in the data set (Miller and Miller, 1998). There are many different methods used to smooth and remove noise from a data set, such as the Fourier transformation or the Savitzky-Golay smoothing filter (Barclay et al. 1997).

Additional processing is typically performed as the first step of PCA or DA analysis. This involves transforming the initial data matrix into a covariance or a correlation matrix. These methods measure how different variables change with respect to other variables. This is the underlying structure that PCA and DA determine similarity or dissimilarity between samples. In general, using a covariance matrix is considered to be without standardization or normalization, as it does not account for the standard deviation of the data. Using correlation (Pearson's) is considered to be a normalization method as the standard deviation is used to scale the data between -1 and +1 (Janzen, 2006). Correlation methods are always applied when the data set being used contains different units (i.e. absorbance and fluorescence data both contained within a data set). There are four main methods of executing these methods: (1) covariance about the origin,

(2) covariance about the mean, (3) correlation about the origin, and (4) correlation about the mean. Rozett and Peterson give a detailed analysis of these four methods and their advantages and disadvantages (Rozett, and McLaughlin Petersen, 1975). In the context of differential sensing, covariance about the origin is the typical approach. Using this pre-processing method prevents the loss of data around the zero point of the experimental scale and avoids the loss of information regarding the relative size and relative error associated with the data from different receptors (Molinowski, 2002).

It is important to note that because pre-processing of data can be key to obtaining the best differentiating model for the data set, some programs that run PCA often include a pre-processing step in the program calculations. In these cases, additional pre-processing may not be necessary for the user and the raw data can be used directly.

# 6 Experimental Setup

Lastly, it is important to note that researchers must take care in their experimental setup to avoid any inherent experimental design flaws that could cause artificial trends. Take for example, microarrays. It has been shown in the literature that if care is not taken with the experimental design one can generate spurious discrimination, where the differences are due to artifacts, such as the days on that the arrays were performed (Chen et al., 2004). Thoughtful care must be taken to ensure that uniform conditions and parameters are applied to each analyte. In addition, when judging the validity of an array, one must also consider the differences between laboratory conditions and real-world

conditions. In order to validate the performance of an array, one must try to replicate real-world conditions and variability.

It is also important to note the difference between technical replicates and experimental replicates. Technical replicates involve replicated data that were derived from using the same stock solutions. These types of replicates help to evaluate pipetting accuracy and the homogeneity of the solutions or media being tested. Experimental replicates require the entire experiment to be reproduced including the growing of cells and preparation of stock solutions. These types of replicates are very useful in preventing results that discriminate data based on irrelevant variables such as the petri dish in that the cells were grown, the well plate in that the array was run, and the conditions in that the solutions or media were stored. Not all systems may require the incorporation of both technical replicates and experimental replicates in a data set, but a clear understanding of the benefits that arise from each different type of replicate may prevent false or unsupported discrimination in a plot, thus avoiding incorrect conclusions.

# 7 Discussion

In this work, the use of statistical analysis tools such as DA and PCA have been discussed in the context of differential sensing. Additionally, a number of key observations regarding the relationships between the data in an array and the corresponding plots have been presented through model data sets. (1) Cross-reactive arrays have demonstrated high discriminatory power and are particularly advantageous over a lock-and-key array when differentiating similar analytes. (2) Optimizing the

number of hosts based on the behavior of an array is an important aspect in designing an array; in particular, the presented examples have emphasized how to choose the best number of hosts for an array and how to recognize circumstances in that adding additional hosts to the array is beneficial. (3) High dimensionality and the benefits and consequences of incorporating high dimensionality into an array were discussed, as well as the importance of investigating the data provided by loading plots and biplots for analyzing receptor performance. (4) The model data sets have shown how to analyze PCA or DA data to obtain the best visual plot representation possible, assuming that visual representation is a goal, and thus learning not to rely exclusively on variance or differentiation as a measure for the quality of an array. (5) The effect of blank or control samples on the appearance of the plot, and the circumstances when an array may not be necessary for differentiating purposes, were explored. (6) Lastly, the implementation of PCA and DA as statistical analysis tools for working up data obtained by sensing arrays was discussed, highlighting the use of validation techniques to probe the effectiveness of the model at representing the data and learning how to avoid bias from experimental design.

## 8 Author contributions

SS Devised experiments and perfomed data analysis and wrote the manuscript. MA assited in experimental design and data analysis. EA overasaw the experiments and provided scientific advice.

# References


Abdi, Hervé. "Factor rotations in factor analyses." *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA* (2003): 792-795.

Adams, Michelle M., and Eric V. Anslyn. "Differential sensing using proteins: exploiting the cross-reactivity of serum albumin to pattern individual terpenes and terpenes in perfume." *Journal of the American Chemical Society* 131.47 (2009): 17068-17069.

Albert, Keith J., et al. "Cross-reactive chemical sensor arrays." *Chemical Reviews* 100.7 (2000): 2595-2626.

Anslyn, Eric V. "Supramolecular analytical chemistry." *The Journal of organic chemistry* 72.3 (2007): 687-699.

Aurup, Helle, David M. Williams, and Fritz Eckstein. "2'-Fluoro and 2-amino-2'-deoxynucleoside 5'-triphosphates as substrates for T7 RNA polymerase."*Biochemistry* 31.40 (1992): 9636-9641.

Bajaj, Avinash et al. “Cell surface-based differentiation of cell types and cancer states using a goldnanoparticle-GFP based sensing array” 1 (2010) 134-138.

Bajaj, Avinash, et al. "Array-based sensing of normal, cancerous, and metastatic cells using conjugated fluorescent polymers." *Journal of the American Chemical Society* 132.3 (2009): 1018-1022.

Bajaj, Avinash, et al. "Detection and differentiation of normal, cancerous, and metastatic cells using nanoparticle-polymer sensor arrays." *Proceedings of the National Academy of Sciences* 106.27 (2009): 10912-10916.

Barclay, V. J., R. F. Bonner, and I. P. Hamilton. "Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression." *Analytical Chemistry* 69.1 (1997): 78-90.

Barfod, Anders, Tina Persson, and Johan Lindh. "In vitro selection of RNA aptamers against a conserved region of the Plasmodium falciparum erythrocyte membrane protein 1." *Parasitology research* 105.6 (2009): 1557-1566.

Barry, Michael J. "Prostate-specific–antigen testing for early diagnosis of prostate cancer." *New England Journal of Medicine* 344.18 (2001): 1373-1377.

Biesecker, Gregory, et al. "Derivation of RNA aptamer inhibitors of human complement C5." *Immunopharmacology* 42.1 (1999): 219-230.

Bratchell, Chemometrics and intelligent laboratory systems 6.2 (1987): 105-125.

Brereton, Richard G., ed. Multivariate pattern recognition in chemometrics: illustrated by case studies. Vol. 9. Elsevier Science, 1992.

Brown, S. Steven D., L. A. Sarabia, and Johan. Trygg. *Comprehensive chemometrics*. Vol. 1. Amsterdam: Elsevier, 2009.

Burke, Donald H., et al. "Bent pseudoknots and novel RNA inhibitors of type 1 human immunodeficiency virus (HIV-1) reverse transcriptase." *Journal of molecular biology* 264.4 (1996): 650-666.

Buryak, Andrey, and Kay Severin. "Dynamic combinatorial libraries of dye complexes as sensors." *Angewandte Chemie* 117.48 (2005): 8149-8152.

Cerchia, Laura, et al. "Differential SELEX in human glioma cell lines." *PloS one*4.11 (2009): e7971

Cerchia, Laura, et al. "Neutralizing aptamers from whole-cell SELEX inhibit the RET receptor tyrosine kinase." *PLoS biology* 3.4 (2005): e123.

Chandrashekar, Jayaram, et al. "The receptors and cells for mammalian taste."*Nature* 444.7117 (2006): 288-294.

Chang, Sam S., et al. "Five different anti-prostate-specific membrane antigen (PSMA) antibodies confirm PSMA expression in tumor-associated neovasculature." *Cancer research* 59.13 (1999): 3192-3198.

Chauveau, Fabien, et al. "Binding of an aptamer to the N-terminal fragment of VCAM-1." *Bioorganic & medicinal chemistry letters* 17.22 (2007): 6119-6122.

Chen, James J., et al. "Analysis of variance components in gene expression data." *Bioinformatics* 20.9 (2004): 1436-1446.

Chu, Ted Chitai. "Anti-cancer and anti-viral aptamers." Dissertation (2006).

Coffin, J. M., Hughes, S. H., Varmus, H. E., Emini, E. A., & Fan, H. Y. (1997). Immunological and Pharmacological Approaches to the Control of Retroviral Infections.

Collins, Byron E., and Eric V. Anslyn. "Pattern-Based Peptide Recognition."*Chemistry-A European Journal* 13.17 (2007): 4700-4708.

Conroy, Paul J., et al. "Antibody production, design and use for biosensor-based applications." *Seminars in cell & developmental biology*. Vol. 20. No. 1. Academic Press, 2009.

Consonni, Viviana, Davide Ballabio, and Roberto Todeschini. "Evaluation of model predictive ability by external validation techniques." *Journal of chemometrics* 24.3-4 (2010): 194-201.

Daniels, Dion A., et al. "Generation of RNA aptamers to the G-protein-coupled receptor for neurotensin, NTS-1." *Analytical biochemistry* 305.2 (2002): 214-226.

Davis, Kenneth A., et al. "Staining of cell surface human CD4 with 2′-F-pyrimidine-containing RNA aptamers for flow cytometry." *Nucleic acids research* 26.17 (1998): 3915-3924.

Dickinson, Todd A., et al. "A chemical-detecting system based on a cross-reactive optical sensor array." *Nature* 382.6593 (1996): 697-700.

Dollins, Claudia M., et al. "Assembling OX40 aptamers on a molecular scaffold to create a receptor-activating aptamer." *Chemistry & biology* 15.7 (2008): 675-682.

Edwards, Nicola Y., et al. "Boronic acid based peptidic receptors for pattern-based saccharide sensing in neutral aqueous media, an application in real-life samples." *Journal of the American Chemical Society* 129.44 (2007): 13575-13583.

Ellington, Andrew D., and Jack W. Szostak. "In vitro selection of RNA molecules that bind specific ligands." *Nature* 346.6287 (1990): 818-822.

Eriksson, L., et al. "Multi-and megavariate data analysis-basic principles and applications, part 1." *Umeå: Umetrics AB* (2006).

Fabrigar, Leandre R., et al. "Evaluating the use of exploratory factor analysis in psychological research." *Psychological methods* 4 (1999): 272-299.

Fitter, Stephen, and Robert James. "Deconvolution of a complex target using DNA aptamers." *Journal of Biological Chemistry* 280.40 (2005): 34193-34201.

Friedman, Jerome H. "Regularized discriminant analysis." *Journal of the American statistical association* 84.405 (1989): 165-175.

Fukunaga, Keinosuke. Introduction to statistical pattern recognition. Academic press, 1990.

Geiger, Albert, et al. "RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity." *Nucleic acids research* 24.6 (1996): 1029-1036.

Gentleman, Robert C., et al. "Bioconductor: open software development for computational biology and bioinformatics." *Genome biology* 5.10 (2004): R80.

Goodey, Adrian, et al. "Development of multianalyte sensor arrays composed of chemically derivatized polymeric microspheres localized in micromachined cavities." *Journal of the American chemical society* 123.11 (2001): 2559-2570.

Haab, Brian B., Maitreya J. Dunham, and Patrick O. Brown. "Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions." *Genome Biol* 2.2 (2001): 1-13.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

Härdle, Wolfgang, and Léopold Simar. *Applied multivariate statistical analysis*. Springer Verlag, 2007.

Hesselberth, Jay R., et al. "In vitro selection of RNA molecules that inhibit the activity of ricin A-chain." *Journal of Biological Chemistry* 275.7 (2000): 4937-4942.

Hicke, Brian J., et al. "Tenascin-C aptamers are generated using tumor cells and purified protein." *Journal of Biological Chemistry* 276.52 (2001): 48644-48654.

Hirsch, Thomas, et al. "A simple strategy for preparation of sensor arrays: molecularly structured monolayers as recognition elements." *Chemical Communications* 3 (2003): 432-433.

Horoszewicz JS, Kawinski E, Murphy GP. Monoclonal antibodies to a new antigenic marker in epithelial prostatic cells and serum of prostatic cancer patients. Anticancer Res 1987;7:927–936.

Hou, Esther W., et al. "High-level expression and purification of untagged and histidine-tagged HIV-1 reverse transcriptase." *Protein expression and purification* 34.1 (2004): 75-86.

Hughes, Andrew D., et al. "A pattern recognition based fluorescence quenching assay for the detection and identification of nitrated explosive analytes."*Chemistry-a European Journal* 14.6 (2008): 1822-1827.

Izenman, Alan Julian. Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, 2008.

James, Tony D., K. R. A. Sandanayake, and Seiji Shinkai. "Saccharide sensing with molecular receptors based on boronic acid." *Angewandte Chemie International Edition in English* 35.17 (1996): 1910-1922.

Janzen, Michael C., et al. "Colorimetric sensor arrays for volatile organic compounds." *Analytical chemistry* 78.11 (2006): 3591-3600.

Jayasena, Sumedha D. "Aptamers: an emerging class of molecules that rival antibodies in diagnostics." *Clinical chemistry* 45.9 (1999): 1628-1650.

Jellinek, D., et al. "High-affinity RNA ligands to basic fibroblast growth factor inhibit receptor binding." *Proceedings of the National Academy of Sciences*90.23 (1993): 11227-11231.

Jeong, Sunjoo, et al. "*In Vitro* Selection of the RNA Aptamer against the Sialyl Lewis X and Its Inhibition of the Cell Adhesion." *Biochemical and biophysical research communications* 281.1 (2001): 237-243.

Jurs, P. C., G. A. Bakken, and H. E. McClelland. "Computational methods for the analysis of chemical sensor array data from volatile analytes." *Chemical Reviews* 100.7 (2000): 2649.

Kaiser, Henry F. "The varimax criterion for analytic rotation in factor analysis."*Psychometrika* 23.3 (1958): 187-200.

Kawasaki, Andrew M., et al. "Uniformly modified 2'-deoxy-2'-fluoro-phosphorothioate oligonucleotides as nuclease-resistant antisense compounds with ,

TB10high affinity and specificity for RNA targets." *Journal of medicinal chemistry* 36.7 (1993): 831-841.

Keefe, Anthony D., Supriya Pai, and Andrew Ellington. "Aptamers as therapeutics." *Nature Reviews Drug Discovery* 9.7 (2010): 537-550.

Kempiak, Stephan J., et al. "Local signaling by the EGF receptor." *Science Signaling* 162.5 (2003): 781.

Kirby, Romy, et al. "Aptamer-based sensor arrays for the detection and quantitation of proteins." *Analytical chemistry* 76.14 (2004): 4066-4075.

Kitamura, Masanori, Shagufta H. Shabbir, and Eric V. Anslyn. "Guidelines for pattern recognition using differential receptors and indicator displacement assays." *The Journal of organic chemistry* 74.12 (2009): 4479-4489.

Klecka, William R. *Discriminant analysis*. No. 19. SAGE Publications, Incorporated, 1980.

Klema, Virginia, and Alan Laub. "The singular value decomposition: Its computation and some applications." *Automatic Control, IEEE Transactions on*25.2 (1980): 164-176.

Korn, Granino Arthur, and Theresa M. Korn. Mathematical handbook for scientists and engineers: definitions, theorems, and formulas for reference and review. Courier Dover Publications, 2000.

Kraus, Elmar, William James, and A. Neil Barclay. "Cutting edge: novel RNA ligands able to bind CD4 antigen and inhibit CD4+ T lymphocyte function." *The Journal of Immunology* 160.11 (1998): 5209-5212.

Lau, Irene PM, et al. "Aptamer-based bio-barcode assay for the detection of cytochrome-*c* released from apoptotic cells." *Biochemical and biophysical research communications* 395.4 (2010): 560-564.

Lavigne, John J., and Eric V. Anslyn. “Sensing A Paradigm Shift in the Field of Molecular Recognition: From Selective to Differential Receptors” *ANGEW CHEM INT EDIT* 40 (2001): 3119-3130.

LeBaron, Matthew J., et al. "Ultrahigh density microarrays of solid samples."*Nature Methods* 2.7 (2005): 511-513.

Lee, Jennifer Fang En.“ Probing aptamer specificity for diagnostics.“ Dissertation (2007).

Lewis, Nathan S. "Comparisons between mammalian and artificial olfaction based on arrays of carbon black-polymer composite vapor detectors." *Accounts of chemical research* 37.9 (2004): 663-672.

Li, Na, et al. "Aptamers that recognize drug-resistant HIV-1 reverse transcriptase." *Nucleic acids research* 36.21 (2008): 6739-6751.

Li, Na, et al. "Directed evolution of gold nanoparticle delivery to cells." *Chem. Commun.* 46.3 (2010): 392-394.

Li, Na, et al. "Inhibition of cell proliferation by an anti-EGFR aptamer." *PloS one*6.6 (2011): e20299.

Li, Yang, et al. "Salivary transcriptome diagnostics for oral cancer detection."*Clinical Cancer Research* 10.24 (2004): 8442-8450.

Li, Yingfu, Ronald Geyer, and Dipankar Sen. "Recognition of anionic porphyrins by DNA aptamers." *Biochemistry* 35.21 (1996): 6911-6922.

Li, Yuan, Hye Jin Lee, and Robert M. Corn. "Detection of protein biomarkers using RNA aptamer microarrays and enzymatically amplified surface plasmon resonance imaging." *Analytical chemistry* 79.3 (2007): 1082-1088.

Long, Stephen B., et al. "Crystal structure of an RNA aptamer bound to thrombin." *Rna* 14.12 (2008): 2504-2512.

Look, Maxime P., et al. "Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients." *Journal of the National Cancer Institute* 94.2 (2002): 116-128.

Lupold, Shawn E., et al. "Identification and characterization of nuclease-stabilized RNA molecules that bind human prostate cancer cells via the prostate-specific membrane antigen." *Cancer Research* 62.14 (2002): 4029-4033.

Luppa, Peter B., Lori J. Sokoll, and Daniel W. Chan. "Immunosensors--principles and applications to clinical chemistry." *Clinica chimica acta; international journal of clinical chemistry* 314.1-2 (2001): 1.

Maddon, Paul Jay, et al. "The isolation and nucleotide sequence of a cDNA encoding the T cell surface protein T4: a new member of the immunoglobulin gene family." *Cell* 42.1 (1985): 93-104.

Madsen, Jeppe B., et al. "RNA aptamers as conformational probes and regulatory agents for plasminogen activator inhibitor-1." *Biochemistry* 49.19 (2010): 4103-4115.

Magalhães, Maria LB, et al. "A general RNA motif for cellular transfection. "*Molecular Therapy* 20.3 (2012): 616-624.

Martinez, Aleix M., and Avinash C. Kak. "Pca versus lda." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.2 (2001): 228-233.

McCauley, Thomas G., Nobuko Hamaguchi, and Martin Stanton. "Aptamer-based biosensor arrays for detection and quantification of biological macromolecules." *Analytical biochemistry* 319.2 (2003): 244-250.

McCleskey, Shawn C., et al. "Differential receptors create patterns diagnostic for ATP and GTP." *Journal of the American Chemical Society* 125.5 (2003): 1114-1115.

McNamara, James O., et al. "Multivalent 4-1BB binding aptamers costimulate CD8+ T cells and inhibit tumor growth in mice." *Journal of Clinical Investigation*118.1 (2008): 376-386.

Meyer, Peter R., et al. "A mechanism of AZT resistance: an increase in nucleotide-dependent primer unblocking by mutant HIV-1 reverse transcriptase."*Molecular cell* 4.1 (1999): 35-43.

Mi, Jing, et al. "Targeted inhibition of αvβ3 integrin with an RNA aptamer impairs endothelial cell growth and survival." *Biochemical and biophysical research communications* 338.2 (2005): 956-963.

Miller, J. C., and J. N. Miller. *Statistics for analytical chemistry, Ellis Harwood Series in Analytical Chemistry, 1988*. ISBN 0-7458-0271-0.

Miller, Jeremy C., et al. "Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers." *Proteomics*3.1 (2003): 56-63.

Miranda, Oscar R. et al., “Array based sensing with nanoparticles: ‘Chemical noses’ for sensing biomolecules and cell surfaces” 14 (2010): 728-736.

Mishra, V. N., and R. P. Agarwal. "Sensitivity, response and recovery time of $SnO_2$ based thick-film sensor array for $H_2$, CO, $CH_4$ and LPG." *Microelectronics Journal* 29.11 (1998): 861-874.

Molinowski, Edmund. "Factor analysis in chemistry." *Wiley, New York 3rd ed.* (2002).

Mor, Gil, et al. "Serum protein markers for early detection of ovarian cancer."*Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005): 7677-7682.

Musto, Christopher J. and Kenneth S. Suslick. “Differential sensing of sugars by colorometric arrays*”* *Current Opinions in Chemical Biology* 14 (2010) 758-766.

Nam, Jwa-Min, Savka I. Stoeva, and Chad A. Mirkin. "Bio-bar-code-based DNA detection with PCR-like sensitivity." *Journal of the American Chemical Society*126.19 (2004): 5932-5933.

Nguyen, Binh T., and Eric V. Anslyn. "Indicator–displacement assays."*Coordination chemistry reviews* 250.23 (2006): 3118-3127.

Ni, Xiaohua, et al. "Nucleic acid aptamers: clinical applications and promising new horizons." *Current medicinal chemistry* 18.27 (2011): 4206.

Niu, Wenze, Nan Jiang, and Yinghe Hu. "Detection of proteins based on amino acid sequences by multiple aptamers against tripeptides." *Analytical biochemistry* 362.1 (2007): 126-135.

Osborne, Scott E., Ichiro Matsumura, and Andrew D. Ellington. "Aptamers as therapeutic and diagnostic reagents: problems and prospects." *Current opinion in chemical biology* 1.1 (1997): 5-9.

Palacios, Manuel A., et al. "Supramolecular chemistry approach to the design of a high-resolution sensor array for multianion detection in water." *Journal of the American Chemical Society* 129.24 (2007): 7538-7544.

Pei, Renjun, et al. “High-resolution cross-reactive array for alkaloids.” *Chem Commun (Camb)* 22 (2009): 3193-3195.

Penazzato Martina, et al.“Effectiveness of antiretroviral therapy in HIV-infected children under 2 years of age.“ Cochrane Database of Systematic Reviews 2012, Issue 7. Art. No.: CD004772. DOI: 10.1002/14651858.CD004772.pub3.

Peng, Gang, et al. "Diagnosing lung cancer in exhaled breath using gold nanoparticles." *Nature nanotechnology* 4.10 (2009): 669-673.

Phillips, Joseph A., et al. "Applications of aptamers in cancer cell biology."*Analytica chimica acta* 621.2 (2008): 101-108.

Proske, Daniela, et al. "Aptamers—basic research, drug development, and clinical applications." *Applied microbiology and biotechnology* 69.4 (2005): 367-374.

Rakow, Neal A., and Kenneth S. Suslick. "A colorimetric sensor array for odour visualization." *Nature* 406.6797 (2000): 710-713.

Rekharsky, Mikhail, et al. "Ion-pairing molecular recognition in water: Aggregation at low concentrations that is entropy-driven." *Journal of the American Chemical Society* 124.50 (2002): 14959-14967.

Risvik, Henning. "Principal Component Analysis (PCA) & NIPALS algorithm." (2007).

Robinson, Dan R., Yi-Mi Wu, and Su-Fang Lin. "The protein tyrosine kinase family of the human genome." *Oncogene* 19.49 (2000): 5548-5557.

Rolland, Olivier, et al. "Dendrimers and nanomedicine: multivalency in action."*New Journal of Chemistry* 33.9 (2009): 1809-1824.

Rougier, Jean-Philippe, et al. "PAI-1 secretion and matrix deposition in human peritoneal mesothelial cell cultures: Transcriptional regulation by TGF-&bgr; 1."*Kidney international* 54.1 (1998): 87-98.

Rozett, Richard W., and E. McLaughlin Petersen. "Methods of factor analysis of mass spectra." *Analytical Chemistry* 47.8 (1975): 1301-1308.

Samuels, Myra L., Jeffrey A. Witmer, and Andrew Schaffner. *Statistics for the life sciences*. Pearson Education, 2012.

Sazani, Peter L., Rosa Larralde, and Jack W. Szostak. "A small aptamer with strong and specific recognition of the triphosphate of ATP." *Journal of the American Chemical Society* 126.27 (2004): 8370-8371.

Sefah, Kwame, et al. "Nucleic acid aptamers for biosensors and bio-analytical applications." *Analyst* 134.9 (2009): 1765-1775.

Sen, Sourav, S. P. Tripathy, and R. S. Paranjape. "Antiretroviral drug resistance testing." *Journal of postgraduate medicine* 52.3 (2006): 187.

Shabbir, Shagufta H., et al. "Pattern-based recognition for the rapid determination of identity, concentration, and enantiomeric excess of subtly different threo diols." *Journal of the American Chemical Society* 131.36 (2009): 13125-13131.

Shangguan, Dihua, et al. "Cell-specific aptamer probes for membrane protein elucidation in cancer cells." *Journal of proteome research* 7.5 (2008): 2133-2139.

Shlens, Jonathon. "A tutorial on principal component analysis." *Systems Neurobiology Laboratory, University of California at San Diego* (2005).

Slinker, B. K., and S. A. Glantz. "Multiple regression for physiological data analysis: the problem of multicollinearity." *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 249.1 (1985): R1-R12.

Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J. Betts, Jaideep Ravela, and Robert W. Shafer "Human immunodeficiency virus reverse transcriptase and protease sequence database." *Nucleic Acids Research,* 31(1) (2003): 298-303.

Spence, Rebecca A., et al. "Mechanism of inhibition of HIV-1 reverse transcriptase by nonnucleoside inhibitors." *Science (New York, NY)* 267.5200 (1995): 988.

Stadtherr, Karin, Hans Wolf, and Petra Lindner. "An aptamer-based protein biochip." *Analytical chemistry* 77.11 (2005): 3437-3443.

Stewart, Sara, et al. "Identifying Protein Variants with Cross-Reactive Aptamer Arrays." *ChemBioChem* 12.13 (2011): 2021-2024.

Stojanovic, Milan N., et al. "Cross-reactive arrays based on three-way junctions." *Journal of the American Chemical Society* 125.20 (2003): 6085-6089.

Stojanovic, Milan N., Paloma De Prada, and Donald W. Landry. "Aptamer-based folding fluorescent sensor for cocaine." *Journal of the American Chemical Society* 123.21 (2001): 4928-4931.

Strehlitz, Beate, Nadia Nikolaus, and Regina Stoltenburg. "Protein detection with aptamer biosensors." *Sensors* 8.7 (2008): 4296-4307.

Suslick, Kenneth S. "An Optoelectronic Nose:"Seeing" Smells by Means of Colorimetric Sensor Arrays" 29 **(**2004**)**: 720-725.

Suslick, Kenneth S., Neal A. Rakow, and Avijit Sen. "Colorimetric sensor arrays for molecular recognition." *Tetrahedron* 60.49 (2004): 11133-11138.

Syrett, Heather Angel. "The application of aptamer microarraying techniques to the detection of HIV-1 reverse transcriptase and its mutant variants." Dissertation (2010).

Taitt, Chris Rowe, et al. "Nine-analyte detection using an array-based biosensor." *Analytical chemistry* 74.23 (2002): 6114-6120.

Takeuchi, Toshihide, et al. "Pattern generation with synthetic sensing systems in lipid bilayer membranes." *Chemical Science* 2.2 (2011): 303-307.

Teixeira, A. M. "Adhesion molecule families: a brief review." (1999).

Theodoridis, Sergios, et al. Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach. Academic Press, 2010.

Tobias, Randall D. "An introduction to partial least squares regression." *Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*. 1995.

Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17.6 (2001): 520-525.

Tuerk, Craig, and Larry Gold. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." *Science*249.4968 (1990): 505-510.

Tuerk, Craig, Sheela MacDougal, and Larry Gold. "RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase."*Proceedings of the National Academy of Sciences* 89.15 (1992): 6988-6992.

Turner, Daniel J., et al. "Toward clinical proteomics on a next-generation sequencing platform." *Analytical chemistry* 83.3 (2010): 666-670.

Ullrich, A., et al. "Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells." *Nature* (1984): 418-425.

Umali, Alona P., and Eric V. Anslyn. "A general approach to differential sensing using synthetic molecular receptors." *Current opinion in chemical biology* 14.6 (2010): 685-692.

Umali, Alona P., et al. "Discrimination of flavonoids and red wine varietals by arrays of differential peptidic sensors." *Chemical Science* 2.3 (2011): 439-445.

Waggoner, Alan. "Fluorescent labels for proteomics and genomics." *Current opinion in chemical biology* 10.1 (2006): 62-66.

Wall, Michael, Andreas Rechtsteiner, and Luis Rocha. "Singular value decomposition and principal component analysis." *A practical approach to microarray data analysis* (2003): 91-109.

Walter, Frank, Quentin Vicens, and Eric Westhof. "Aminoglycoside–RNA interactions." *Current opinion in chemical biology* 3.6 (1999): 694-704.

Wassermann, Eric M., et al. "Noninvasive mapping of muscle representations in human motor cortex." *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 85.1 (1992): 1-8.

Wehrens, Ron, Hein Putter, and Lutgarde Buydens. "The bootstrap: a tutorial."*Chemometrics and Intelligent Laboratory Systems* 54.1 (2000): 35-52.

Weiss, Tristen C., et al. "An RNA aptamer with high affinity and broad specificity for zinc finger proteins." *Biochemistry* 49.12 (2010): 2732-2740.

Wold, Herman. *Nonlinear Iterative Partial Least Squares (NIPLAS) Modelling: Some Current Developments*. Univ., Department of Statistics, 1973.

Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2.1 (1987): 37-52.

Wright, Aaron T., et al. "The Discriminatory Power of Differential Receptor Arrays Is Improved by Prescreening—A Demonstration in the Analysis of Tachykinins and Similar Peptides" *ANGEW CHEM INT EDIT* 46 *(*2007): 8212-8215.

Xiao, Zeyu, et al. "Cell-Specific Internalization Study of an Aptamer from Whole Cell Selection." *Chemistry-A European Journal* 14.6 (2008): 1769-1775.

Yeh, Edward TH, and James T. Willerson. "Coming of Age of C-reactive protein using Inflammation markers in cardiology." *Circulation* 107.3 (2003): 370-371.

Zhang, Chen, and Kenneth S. Suslick. "A colorimetric sensor array for organics in water." *Journal of the American Chemical Society* 127.33 (2005): 11548-11549.

Zhang, Chen, and Kenneth S. Suslick. "Colorimetric sensor array for soft drink analysis." *Journal of agricultural and food chemistry* 55.2 (2007): 237-242.

Zhang, Hongtao, et al. "ErbB receptors: from oncogenes to targeted cancer therapies." *Journal of Clinical Investigation* 117.8 (2007): 2051-2058.

Zhang, Chen, Daniel P. Bailey, and Kenneth S. Suslick. "Colorimetric sensor arrays for the analysis of beers: a feasibility study." *Journal of agricultural and food chemistry* 54.14 (2006): 4925-4931.

Zhang, Xiao-Bing, Rong-Mei Kong, and Yi Lu. "Metal ion sensors based on DNAzymes and related DNA molecules." *Annual review of analytical chemistry (Palo Alto, Calif.)* 4.1 (2011): 105.

Zhou, Huchen, et al. "Pattern recognition of proteins based on an array of functionalized porphyrins." *Journal of the American Chemical Society* 128.7 (2006): 2421-2425.

# Vita

Sara Stewart was born October 21st, 1980 in Atlanta, GA. In 1994 she moved to Mesa, AZ where she attended Mt. View High School and attained advanced placement in biology and chemistry. She began her academic career at Arizona State University in 1998 where she studied chemical engineering. In 1999 she left Arizona State University to become better prepared for advanced studies and took general classes at Mesa Community College from 1999-2002. It was here that she developed her fascination with biology and the intricate working of life. In 2002 she returned to Arizona State University to study Molecular Bioscience and Biotechnology (MBB). She graduated with a B.S. in Molecular Bioscience and Biotechnology with a minor in Anthropology as part of the MBB programs third graduating class. Between 2001 and 2004 she worked as a laboratory supervisor at ZLB plasma services; between 2004 and 2006 she performed consumer clinical research for Hill-Top Inc. In 2006 she began her graduate career as a member of the McDevitt lab where she researched antibody microarrays for cardiovascular prognostics. In 2010 she joined the Anslyn and Ellington labs to continue

research on arrays and shifted her focused to aptamers. Sara Stewart had her first child, Garrett Goodwin in February 2006 and married her husband Tracy Goodwin in April 2006. Her daughter Brianna Goodwin was born in February 2008. Sara is currently employed as the Manager of Technology Development in the McCombie lab at Cold Spring Harbor Laboratory. She can be contacted at sgoodwin@cshl.edu.

This dissertation was typed by the author: Sara Stewart.